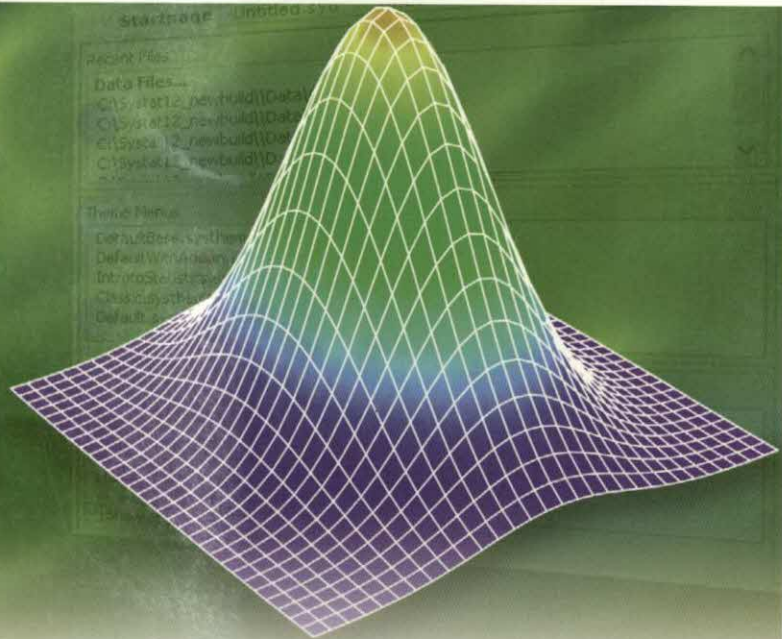


SYSTAT[®] 12



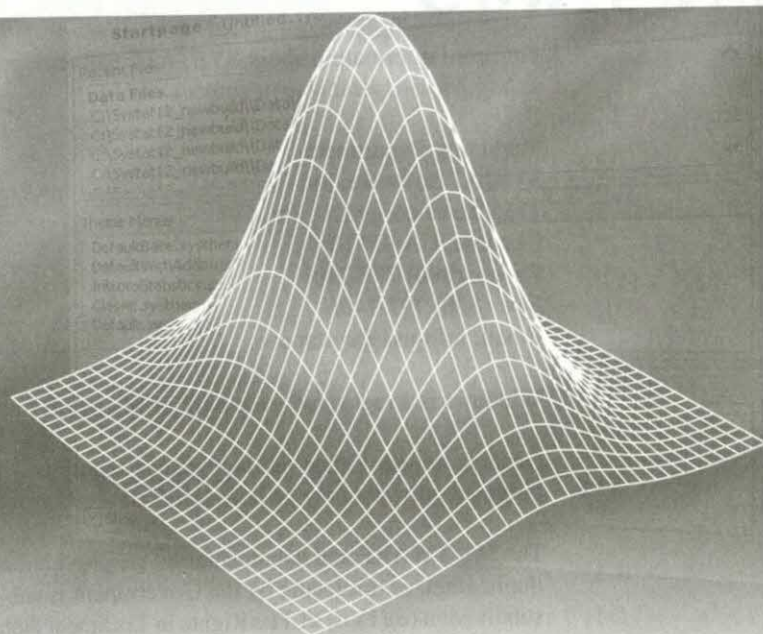
Statistics II

~~Ref~~
~~Lib~~
PS
2/18

For Acc
SCERT library

PS
28/10/10

SYSTAT[®] 12



Statistics II



SYSTAT
WWW.SYSTAT.COM

For more information about SYSTAT[®] software products, please visit our WWW site at <http://www.systat.com> or contact

Marketing Department
SYSTAT Software, Inc.
1735 Technology Dr., Ste. 430
San Jose, CA 95110
Phone: (800) 797-7401
Fax: (800) 797-7406
Email: info-usa@systat.com

Windows is a registered trademark of Microsoft Corporation.

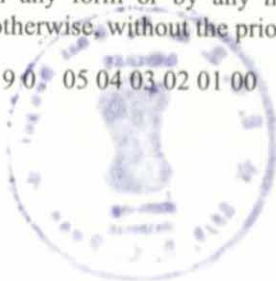
General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c)(1)(ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacture is SYSTAT Software, Inc., 1735 Technology Drive, Suite 430, San Jose, CA 95110. USA.

SYSTAT[®] 12 Statistics- II
Copyright © 2007 by SYSTAT Software, Inc.
SYSTAT Software, Inc.
1735 Technology Dr., Ste. 430
San Jose, CA 95110
All rights reserved.
Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 0 05 04 03 02 01 00



1.06.2010
13999

Contents

List of Examples

xxxiii

Statistics I

1 Introduction to Statistics

I-1

Descriptive Statistics	I-1
Know Your Batch	I-2
Sum, Mean, and Standard Deviation	I-3
Stem-and-Leaf Plots	I-3
The Median	I-4
Sorting	I-5
Standardizing	I-6
Inferential Statistics.	I-7
What is a Population?	I-7
Picking a Simple Random Sample.	I-8
Specifying a Model	I-10
Estimating a Model	I-10
Confidence Intervals.	I-11
Hypothesis Testing.	I-12
Checking Assumptions	I-14
References	I-16

2 Bootstrapping and Sampling

I-17

Statistical Background	I-17
Resampling in SYSTAT	I-21
Resampling Tab.	I-21
Using Commands	I-22
Usage Considerations	I-22
Examples.	I-23
Computation	I-38
Algorithms	I-38
Missing Data	I-38
References	I-39

3 Classification and Regression Trees

I-41

Statistical Background	I-42
The Basic Tree Model.	I-42
Categorical or Quantitative Predictors	I-45
Regression Trees	I-45
Classification Trees	I-46
Stopping Rules, Pruning, and Cross-Validation	I-47
Loss Functions	I-48
Geometry	I-48
Classification and Regression Trees in SYSTAT	I-51
Classification and Regression Trees Dialog Box	I-51
Using Commands	I-54
Usage Considerations	I-54
Examples.	I-54
Computation	I-62
Algorithms	I-62
Missing Data	I-62
References	I-62

4 Cluster Analysis

I-65

Statistical Background.	I-66
Types of Clustering.	I-66
Correlations and Distances.	I-67
Hierarchical Clustering.	I-68
Partitioning via K-Clustering.	I-78
Additive Trees.	I-80
Cluster Analysis in SYSTAT.	I-82
Hierarchical Clustering Dialog Box.	I-82
K-Clustering Dialog Box.	I-88
Additive Trees Clustering Dialog Box.	I-91
Using Commands.	I-93
Usage Considerations.	I-95
Examples.	I-96
Computation.	I-122
Algorithms.	I-122
Missing Data.	I-122
References.	I-122

5 Conjoint Analysis

I-125

Statistical Background.	I-125
Additive Tables.	I-126
Multiplicative Tables.	I-128
Computing Table Margins Based on an Additive Model.	I-130
Applied Conjoint Analysis.	I-131
Conjoint Analysis in SYSTAT.	I-133
Conjoint Analysis Dialog Box.	I-133
Using Commands.	I-135
Usage Considerations.	I-135
Examples.	I-136

Computation	I-152
Algorithms	I-152
Missing Data	I-153
References	I-153

6 Correlations, Associations, and Distance Measures ***I-157***

Statistical Background	I-158
The Scatterplot Matrix (SPLOM).	I-159
The Pearson Correlation Coefficient	I-160
Other Measures of Association	I-161
Transposed Data	I-167
Hadi Robust Outlier Detection	I-168
Simple Correlations in SYSTAT	I-170
Simple Correlations Dialog Box	I-170
Using Commands	I-177
Usage Considerations	I-178
Examples.	I-179
Computation	I-199
Algorithms	I-199
Missing Data	I-199
References	I-200

7 Correspondence Analysis ***I-201***

Statistical Background	I-201
The Simple Model	I-202
The Multiple Model.	I-203
Correspondence Analysis in SYSTAT	I-204
Correspondence Analysis Dialog Box	I-204

Smart Correspondence Analysis Dialog Box.	I-205
Using Commands.	I-206
Usage Considerations.	I-206
Examples	I-207
Computation.	I-218
Algorithms	I-218
Missing Data	I-218
References	I-218

8 Crosstabulation

(One-Way, Two-Way, and Multiway) **I-219**

Statistical Background.	I-220
Making Tables	I-220
Significance Tests and Measures of Association.	I-222
Crosstabulations in SYSTAT	I-228
One-Way Frequency Tables Dialog Box.	I-228
Two-Way Tables Dialog Box	I-231
Multiway Tables: Tabulate Dialog Box	I-237
Using Commands.	I-244
Usage Considerations.	I-246
Examples	I-248
References	I-296

9 Descriptive Statistics

I-297

Statistical Background.	I-299
Location.	I-299
Spread.	I-301
The Normal Distribution	I-301
Test for Normality	I-302

Multivariate Normality Assessment	I-303
Non-Normal Shape	I-303
Subpopulations	I-305
Descriptive Statistics in SYSTAT	I-307
Basic Statistics Dialog Box	I-307
Stem-and-Leaf Plot Dialog Box	I-314
Basic Statistics for Rows	I-316
Row Stem-and-Leaf Plot Dialog Box	I-320
Cronbach's Alpha Dialog Box	I-321
Using Commands	I-322
Usage Considerations	I-323
Examples	I-324
Computation	I-344
Algorithms	I-344
References	I-344

10 Design of Experiments

I-345

Statistical Background	I-346
The Research Problem	I-346
Types of Investigation	I-347
The Importance of Having a Strategy	I-348
The Role of Experimental Design in Research	I-349
Types of Experimental Designs	I-349
Factorial Designs	I-350
Response Surface Designs	I-354
Mixture Designs	I-357
Optimal Designs	I-362
Choosing a Design	I-366
Design of Experiments in SYSTAT	I-368
Design of Experiments Wizard	I-368
Classic Design of Experiments	I-369
Using Commands	I-370

Usage Considerations	I-370
Examples	I-371
References	I-388

11 Discriminant Analysis

I-391

Statistical Background	I-392
Linear Discriminant Model	I-392
Robust Discriminant Analysis	I-399
Discriminant Analysis in SYSTAT	I-400
Classical Discriminant Analysis Dialog box	I-400
Robust Discriminant Analysis Dialog Box	I-405
Using Commands	I-407
Usage Considerations	I-408
Examples	I-409
References	I-450

12 Factor Analysis

I-453

Statistical Background	I-453
A Principal Component	I-454
Factor Analysis	I-457
Principal Components versus Factor Analysis	I-460
Applications and Caveats	I-461
Factor Analysis in SYSTAT	I-462
Factor Analysis Dialog Box	I-462
Using Commands	I-468
Usage Considerations	I-468
Examples	I-469
Computation	I-492
Algorithms	I-492

Missing Data	I-492
References	I-493

13 Fitting Distributions ***I-495***

Statistical Background	I-495
Goodness-of-Fit Tests.	I-496
Fitting Distributions in SYSTAT	I-498
Fitting Distributions: Discrete Dialog Box	I-498
Fitting Distributions: Continuous Dialog Box	I-499
Using Commands	I-501
Usage Considerations	I-503
Examples.	I-503
Computation	I-518
Algorithms	I-518
References	I-518

14 Hypothesis Testing ***I-519***

Statistical Background	I-520
One-Sample Tests and Confidence Intervals for Mean and Proportion	I-520
Two-Sample Tests and Confidence Intervals for Means and Proportions	I-520
Tests for Variances and Confidence Intervals	I-521
Tests for Correlations and Confidence Intervals	I-522
Multiple Tests	I-522
Hypothesis Testing in SYSTAT	I-523
Tests for Mean(s)	I-523
Tests for Variance(s)	I-531
Tests for Correlation(s)	I-535
Tests for Proportion(s)	I-538

Using Commands	I-541
Usage Considerations	I-543
Examples	I-544
References	I-566

Statistics II

1 Linear Models

II-1

Simple Linear Models	II-1
Equation for a Line	II-2
Least Squares	II-5
Estimation and Inference	II-5
Standard Errors	II-7
Hypothesis Testing	II-7
Multiple Correlation	II-8
Regression Diagnostics	II-9
Multiple Regression	II-12
Variable Selection	II-15
Using an SSCP, a Covariance, or a Correlation Matrix as Input	II-18
Analysis of Variance	II-19
Effects Coding	II-20
Means Coding	II-21
Models	II-22
Hypotheses	II-23
Multigroup ANOVA	II-24
Factorial ANOVA	II-24
Data Screening and Assumptions	II-25
Levene Test	II-25
Pairwise Mean Comparisons	II-26

Linear and Quadratic Contrasts	II-28
Repeated Measures	II-31
Assumptions in Repeated Measures	II-32
Issues in Repeated Measures Analysis	II-33
SYSTAT's Sum of Squares	II-34
References	II-36

2 Linear Models I: Linear Regression II-39

Linear Regression in SYSTAT	II-41
Least Squares Regression Dialog Box	II-41
Ridge Regression	II-48
Ridge Regression Dialog Box.	II-49
Bayesian Regression	II-50
Bayesian Regression Dialog Box	II-51
Using Commands	II-53
Usage Considerations	II-54
Examples.	II-55
Computation	II-104
Algorithms	II-104
References	II-104

3 Linear Models II: Analysis of Variance II-107

Analysis of Variance in SYSTAT	II-108
Analysis of Variance: Estimate Model Dialog Box.	II-108
Analysis of Variance: Hypothesis Test Dialog Box	II-113
Analysis of Variance: Pairwise Comparisons Dialog Box	II-117
Using Commands	II-121
Usage Considerations	II-121
Examples.	II-122

ComputationII-171
AlgorithmsII-171
ReferencesII-171

4 Linear Models III:General Linear ModelsII-175

General Linear Models in SYSTAT.II-177
Model Estimation (in GLM)II-177
Hypothesis Test.II-186
Pairwise ComparisonsII-195
Post hoc Tests for Repeated MeasuresII-199
Using Commands.II-200
Usage Considerations.II-201
ExamplesII-203
ComputationII-249
AlgorithmsII-249
ReferencesII-249

5 Introduction to Linear Mixed Models II-251

Mixed Models and Paired t-testII-251
Fixed Effects Versus Random EffectsII-255
Why Use Random Effects?II-259
Some Linear Model TerminologyII-261
String and Numeric VariablesII-261
EstimabilityII-262
Data Layout: Multiway or NestedII-262
Nested LayoutII-266
Balanced and Unbalanced Data.II-267
SYSTAT Notation for Random EffectsII-267
Covariance StructuresII-269

Using Covariates: Regression	II-276
Estimation and Prediction	II-279
Estimating the Fixed Effects	II-279
Estimating Covariance Matrices	II-281
Testing Hypotheses	II-286
The F Matrix	II-287
The D Matrix	II-288
The R Matrix	II-289
Pairwise Comparison Tests	II-290
Diagnostics	II-290
Residual Diagnostics	II-291
Further Insights	
Henderson's Mixed Model EquationII-293	
Some Properties of BLUPs	II-294
Why Random Effect Coefficients are Always Estimable. . .	II-295
ML and REML	II-295
References	II-297

6 Variance Components Models **II-299**

Statistical Background	II-299
Variance Components in SYSTAT	II-301
Model Estimation (in VC)	II-301
Hypothesis Test	II-306
Using Commands	II-310
Usage Considerations	II-310
Examples	II-311
References	II-342

7 Linear Mixed Models **II-343**

Statistical Background	II-344
----------------------------------	--------

Linear Mixed Models in SYSTAT	II-345
Model Estimation (in MIXED).	II-345
Category	II-347
Random	II-348
Options	II-350
Hypothesis Tests	II-352
F and R Matrices	II-354
D Matrix	II-355
Using Commands.	II-356
Usage Considerations	II-356
Examples	II-357
References	II-384

8 Hierarchical Linear Mixed Models **II-385**

Statistical Background.	II-386
Hierarchical Linear Mixed Models in SYSTAT	II-387
Model Estimation (in MIXED).	II-387
Hypothesis Test.	II-394
Using Commands.	II-398
Usage Considerations.	II-398
Examples	II-399
References	II-419

9 Mixed Regression **II-421**

Statistical Background.	II-422
Historical Approaches	II-423
The General Mixed Regression Model.	II-424
Model Comparisons	II-431
Mixed Regression in SYSTAT	II-431

Mixed Regression: Hierarchical Data.	II-431
Data Structure.	II-438
Using Commands.	II-441
Usage Considerations.	II-441
Examples.	II-442
Computation.	II-484
Algorithms.	II-484
References.	II-485

Statistics III

1 Logistic Regression

III-1

Statistical Background.	III-2
Binary Logit.	III-2
Multinomial Logit.	III-5
Conditional Logit.	III-5
Discrete Choice Logit.	III-7
Stepwise Logit.	III-9
Logistic Regression in SYSTAT.	III-10
Estimate Model Dialog Box.	III-10
Quantiles.	III-18
Simulation.	III-19
Hypothesis.	III-20
Using Commands.	III-22
Usage Considerations.	III-22
Examples.	III-24
Computation.	III-85
Algorithms.	III-85
Missing Data.	III-86

References	III-89
----------------------	--------

2 Loglinear Models ***III-93***

Statistical Background.	III-94
Fitting a Loglinear Model	III-95
Loglinear Models in SYSTAT	III-96
Loglinear Model: Estimate Dialog Box	III-96
Frequency Table (Tabulate)	III-102
Using Commands	III-103
Usage Considerations.	III-103
Examples	III-105
Computation.	III-122
Algorithms	III-122
References	III-122

3 Missing Value Analysis ***III-123***

Statistical Background.	III-123
Techniques for Handling Missing Values	III-125
Randomness and Missing Data	III-131
Testing for Randomness	III-133
A Final Caution.	III-134
Missing Value Analysis in SYSTAT	III-134
Missing Value Analysis Dialog Box	III-134
Using Commands	III-136
Usage Considerations.	III-137
Examples	III-137
Computation.	III-183
Algorithms	III-183
References	III-184

4 Multidimensional Scaling

III-185

Statistical Background	III-186
Assumptions.	III-186
Collecting Dissimilarity Data	III-187
Scaling Dissimilarities	III-188
Multidimensional Scaling in SYSTAT	III-189
Multidimensional Scaling Dialog Box	III-189
Using Commands	III-194
Usage Considerations	III-194
Examples.	III-195
Computation	III-210
Algorithms	III-211
Missing Data	III-212
References	III-213

5 Multinormal Tests

III-215

Statistical Background	III-215
Multinormal Tests in SYSTAT	III-216
Multinormal Tests Dialog Box	III-216
Using Commands	III-217
Usage Considerations	III-217
Examples.	III-218
References	III-221

6 Multivariate Analysis of Variance

III-223

Statistical Background	III-224
MANOVA Tests	III-225
MANOVA in SYSTAT	III-227

MANOVA: Estimate Model Dialog Box	III-227
Hypothesis Test Dialog Box	III-232
Between-Groups Testing	III-239
Within-Group Testing	III-241
Post hoc Test for Repeated measures.	III-242
Using Commands	III-244
Usage Considerations.	III-244
Examples	III-246
References	III-259

7 Nonlinear Models

III-261

Statistical Background.	III-262
Modeling the Dose-Response Function	III-262
Loss Functions	III-265
Model Estimation.	III-269
Problems	III-269
Nonlinear Models in SYSTAT	III-270
Nonlinear Regression: Estimate Model	III-270
Loss Functions for Analytic Function Minimization.	III-281
Using Commands	III-283
Usage Considerations.	III-283
Examples	III-284
Computation	III-316
Algorithms	III-316
Missing Data	III-316
References	III-318

8 Nonparametric Tests

III-319

Statistical Background.	III-320
---------------------------------	---------

Rank (Ordinal) Data.	III-320
Categorical (Nominal) Data.	III-321
Robustness	III-321
Nonparametric Tests for Independent Samples in SYSTAT . . .	III-322
Kruskal-Wallis Test Dialog Box	III-322
Two-Sample Kolmogorov-Smirnov Test Dialog Box	III-323
Using Commands	III-325
Nonparametric Tests for Related Variables in SYSTAT	III-325
Sign Test Dialog Box	III-325
Wilcoxon Signed-Rank Test Dialog Box	III-326
Friedman Test Dialog Box	III-328
Quade Test Dialog Box	III-329
Using Commands	III-331
Nonparametric Tests for Single Samples in SYSTAT	III-331
One-Sample Kolmogorov-Smirnov Test Dialog Box	III-331
Anderson-Darling Test Dialog Box.	III-334
Wald-Wolfowitz Runs Test Dialog Box	III-337
Using Commands	III-338
Usage Considerations	III-339
Examples.	III-340
Computation	III-355
Algorithms	III-355
References	III-355

9 Partial Least Squares Regression III-357

Statistical Background.	III-357
Model Building	III-358
Cross-Validation	III-360
Partial Least Squares Regression in SYSTAT.	III-361
Partial Least Squares Regression Dialog Box	III-361
Using Commands	III-364
Usage Considerations	III-364

Examples	III-365
Computation	III-377
Algorithms	III-377
Missing Data	III-378
References	III-378

10 Partially Ordered Scalogram Analysis with Coordinates ***III-381***

Statistical Background.	III-381
Coordinates	III-383
POSAC in SYSTAT.	III-384
POSAC Dialog Box	III-384
Using Commands.	III-385
Usage Considerations.	III-385
Examples	III-386
Computation	III-395
Algorithms	III-395
Missing Data	III-395
References	III-395

11 Path Analysis (RAMONA) ***III-397***

Statistical Background.	III-397
The Path Diagram.	III-397
Path Analysis in SYSTAT.	III-405
Instructions for using RAMONA.	III-405
The MODEL statement.	III-407
RAMONA Options	III-411
Usage Considerations.	III-413
Examples	III-414

Computation	III-452
RAMONA's Model	III-452
Algorithms	III-454
References	III-460
Acknowledgments	III-461

Statistics IV

1 Perceptual Mapping

IV-1

Statistical Background	IV-1
Preference Mapping	IV-2
Biplots and MDPREF	IV-6
Procrustes Rotations	IV-7
Perceptual Mapping in SYSTAT	IV-7
Perceptual Mapping Dialog Box	IV-7
Using Commands	IV-9
Usage Considerations	IV-9
Examples	IV-9
Computation	IV-16
Algorithms	IV-16
Missing data	IV-16
References	IV-16

2 Power Analysis

IV-19

Statistical Background	IV-20
Error Types	IV-21
Power	IV-22

Displaying Power Results	IV-32
Generic Power Analysis	IV-34
Power Analysis in SYSTAT.	IV-39
Single Proportion	IV-39
Equality of Two Proportions	IV-40
Single Correlation Coefficient	IV-42
Equality of Two Correlation Coefficients	IV-44
One-Sample z-test	IV-46
Two-Sample z-test	IV-48
One-Sample t-test.	IV-50
Paired t-test	IV-51
Two-Sample t-test	IV-53
One-Way ANOVA	IV-55
Two-Way ANOVA.	IV-57
Generic Power Analysis	IV-60
Using Commands.	IV-62
Usage Considerations.	IV-62
Examples	IV-63
Computation.	IV-83
Algorithms	IV-83
References	IV-83

3 Probability Calculator

IV-85

Statistical Background.	IV-85
Probability Calculator in SYSTAT	IV-86
Univariate Discrete Distributions Dialog Box	IV-86
Univariate Continuous Distributions Dialog Box	IV-87
Using Commands.	IV-90
Usage Considerations.	IV-90
Examples	IV-90
References	IV-98

4 Probit Analysis

IV-99

Statistical Background	IV-99
Interpreting the Results	IV-100
Probit Analysis in SYSTAT	IV-100
Probit Regression Dialog Box	IV-100
Using Commands	IV-103
Usage Considerations	IV-103
Examples	IV-104
Computation	IV-107
Algorithms	IV-107
Missing Data	IV-107
References	IV-107

5 Quality Analysis

IV-109

Statistical Background	IV-109
Quality Analysis in SYSTAT	IV-110
Histogram	IV-110
Quality Analysis: Histogram Dialog Box	IV-110
Pareto Charts	IV-111
Pareto Chart Dialog Box	IV-112
Box-and-Whisker Plots	IV-112
Box-and-Whisker Plot Dialog Box	IV-113
Control Charts	IV-114
Run Charts	IV-114
Run Chart Dialog Box	IV-115
Shewhart Control Charts	IV-116
Shewhart Control Chart Dialog Box	IV-116
OC and ARL curves	IV-134
Operating Characteristic Curves	IV-135
Operating Characteristic Curve Dialog Box	IV-135
Average Run Length Curves	IV-136

Average Run Length Dialog Box	IV-137
Cusum Charts	IV-142
Cumulative Sum Chart Dialog Box	IV-142
Moving Average Charts	IV-144
Moving Average Chart Dialog Box	IV-144
Exponentially Weighted Moving Average Charts	IV-146
Exponentially Weighted Moving Average Chart Dialog Box	IV-146
X-MR Charts	IV-149
X-MR Chart Dialog Box	IV-150
Regression Charts.	IV-152
Regression Chart Dialog Box.	IV-152
TSQ Charts	IV-153
TSQ Chart Dialog Box	IV-154
Process Capability Analysis	IV-155
Process Capability Analysis Dialog Box	IV-159
Using Commands.	IV-161
Usage Considerations.	IV-162
Examples	IV-163
References	IV-217

6 Random Sampling

IV-219

Statistical Background.	IV-220
Random Sampling in SYSTAT	IV-220
Univariate Discrete Distributions Dialog Box	IV-220
Univariate Continuous Distributions Dialog Box	IV-222
Using Commands.	IV-223
Distribution Notations used in Random Sampling	IV-223
Usage Considerations.	IV-224
Examples	IV-225
Computation	IV-228
Algorithms	IV-228
References	IV-228

7 Response Surface Methods **IV-231**

Statistical Background	IV-231
Fitting a Response Surface	IV-232
Contour and Surface plot	IV-233
Response Optimization	IV-234
Response Surface Methods in SYSTAT	IV-237
Response Surface Methods: Optimize Dialog Box	IV-240
Using Commands	IV-244
Usage Considerations	IV-244
Examples	IV-245
Computation	IV-252
References	IV-253

8 Robust Regression **IV-255**

Statistical Background	IV-256
Least Absolute Deviations (LAD) Regression	IV-260
M Regression	IV-261
Least Median Squares (LMS) Regression	IV-261
Least Trimmed Squares (LTS) Regression	IV-261
Scale (S) Regression	IV-262
Rank Regression	IV-262
Asymptotic Standard Errors, Confidence Intervals and Robust R2	IV-262
Robust Regression in SYSTAT	IV-263
Least Absolute Deviation (LAD) Regression Dialog Box . .	IV-263
M Regression Dialog Box	IV-265
Least Median of Squares (LMS) Regression Dialog Box . .	IV-268
Least Trimmed Squares (LTS) Regression Dialog Box . . .	IV-271
S Regression Dialog Box	IV-275
Rank Regression Dialog Box	IV-278
Using Commands	IV-279
Usage Considerations	IV-279

Examples	IV-280
Computation	IV-287
Algorithms	IV-287
Missing Data	IV-288
References	IV-288

9 Set and Canonical Correlations IV-291

Statistical Background.	IV-291
Sets	IV-292
Partialing	IV-292
Notation.	IV-293
Measures of Association Between Sets.	IV-293
$R^2_{Y,X}$ Proportion of Generalized Variance	IV-293
$T^2_{Y,X}$ and $P^2_{Y,X}$ Proportions of Additive Variance	IV-294
Interpretations.	IV-295
Types of Association between Sets.	IV-296
Testing the Null Hypothesis	IV-297
Estimates of the Population $R^2_{Y,X}$, $T^2_{Y,X}$, and $P^2_{Y,X}$	IV-299
Set and Canonical Correlations in SYSTAT	IV-299
Set and Canonical Correlations Dialog Box	IV-299
Category	IV-301
Options	IV-303
Using Commands.	IV-304
Usage Considerations.	IV-304
Examples	IV-305
Computation	IV-315
Algorithms	IV-315
Missing Data	IV-316
References	IV-316

10 Signal Detection Analysis

IV-319

Statistical Background	IV-319
Detection Parameters	IV-320
Signal Detection Analysis in SYSTAT	IV-321
Signal Detection Analysis Dialog Box	IV-321
Using Commands	IV-324
Usage Considerations	IV-325
Examples	IV-328
Computation	IV-346
Algorithms	IV-346
Missing Data	IV-346
References	IV-346

11 Smoothing

IV-349

Statistical Background	IV-350
The Three Ingredients of Nonparametric Smoothers	IV-350
A Sample Data Set	IV-351
Kernels	IV-352
Bandwidth	IV-355
Smoothing Functions	IV-358
Smoothness	IV-360
Interpolation and Extrapolation	IV-360
Close Relatives (Roses by Other Names)	IV-360
Smoothing in SYSTAT	IV-362
Smooth & Plot Dialog Box	IV-362
Using Commands	IV-366
Usage Considerations	IV-366
Examples	IV-367
References	IV-382

12 Spatial Statistics

IV-385

Statistical Background.	IV-385
The Basic Spatial Model	IV-385
The Geostatistical Model	IV-387
Variogram.	IV-388
Variogram Models	IV-389
Anisotropy	IV-392
Simple Kriging	IV-393
Ordinary Kriging	IV-394
Universal Kriging.	IV-394
Simulation	IV-394
Point Processes	IV-395
Spatial Statistics in SYSTAT	IV-399
Spatial Statistics Dialog Box	IV-399
Using Commands.	IV-408
Usage Considerations.	IV-410
Examples	IV-411
Computation.	IV-426
Missing Data	IV-426
Algorithms	IV-426
References	IV-426

13 Survival Analysis

IV-427

Statistical Background.	IV-428
Graphics	IV-429
Parametric Modeling	IV-432
Survival Analysis in SYSTAT	IV-435
Survival Analysis: Nonparametric Dialog Box.	IV-436
Survival Analysis: Parametric and Cox Dialog Box	IV-439
Using Commands.	IV-447

Usage Considerations	IV-448
Examples.	IV-449
Computation	IV-476
Algorithms	IV-476
Missing Data	IV-476
References	IV-484

14 Test Item Analysis

IV-487

Statistical Background	IV-488
Classical Model	IV-489
Latent Trait Model	IV-490
Test Item Analysis in SYSTAT	IV-491
Classical Test Item Analysis Dialog Box	IV-491
Logistic Test Item Analysis Dialog Box	IV-493
Using Commands	IV-494
Usage Considerations	IV-495
Examples.	IV-498
Computation	IV-506
Algorithms	IV-506
Missing Data	IV-507
References	IV-507

15 Time Series

IV-509

Statistical Background	IV-510
Smoothing.	IV-510
ARIMA Modeling and Forecasting.	IV-514
Seasonal Decomposition and Adjustment	IV-523
Exponential Smoothing	IV-524
Trend Analysis	IV-525

Fourier Analysis	IV-526
Graphical Displays for Time Series in SYSTAT	IV-528
Time Axis Format Dialog Box	IV-528
Time Series Plot Dialog Box	IV-529
ACF Plot Dialog Box	IV-529
PACF Plot Dialog Box	IV-530
CCF Plot Dialog Box	IV-531
Using Commands	IV-532
Transformations of Time Series in SYSTAT	IV-532
Transform Dialog Box	IV-532
Clear Series	IV-534
Using Commands	IV-534
Smoothing a Time Series in SYSTAT	IV-535
Moving Average Smoothing Dialog Box	IV-535
LOWESS Smoothing Dialog Box	IV-536
Exponential Smoothing Dialog Box	IV-537
Using Commands	IV-539
Seasonal Adjustments in SYSTAT	IV-539
Seasonal Adjustment Dialog Box	IV-539
Using Commands	IV-540
ARIMA Models in SYSTAT	IV-540
ARIMA Dialog Box	IV-540
Using Commands	IV-542
Trend Analysis in SYSTAT	IV-542
Trend Analysis dialog box	IV-542
Using Commands	IV-544
Fourier Models in SYSTAT	IV-544
Fourier Transformation Dialog Box	IV-545
Using Commands	IV-546
Usage Considerations	IV-546
Examples	IV-547
Computation	IV-578
Algorithms	IV-578
References	IV-578

16 Two-Stage Least Squares IV-581

Statistical Background	IV-581
Two-Stage Least Squares Estimation	IV-582
Heteroskedasticity	IV-583
Two-Stage Least Squares in SYSTAT	IV-584
Two-Stage Least Squares Regression Dialog Box	IV-584
Using Commands	IV-586
Usage Considerations	IV-586
Examples	IV-587
Computation	IV-597
Algorithms	IV-597
Missing Data	IV-597
References	IV-597

Acronym & Abbreviation Expansions

Index

List of Examples

Multi Way: Standardize Tables	I-291
A Model with Interaction	II-315
A Nested-Factorial Model with Case Frequencies	II-412
Actuarial Life Tables	IV-453
Additive Trees	I-120
AIC and Schwarz's BIC	III-258
Analysis of Covariance (ANCOVA)	II-209
Analysis of Covariance	II-153
Anderson-Darling Test	III-353
ANOVA Assumptions and Contrasts	II-126
ARIMA Models	IV-566
ARL Curve	IV-197
Autocorrelation Plot	IV-548
Automatic Stepwise Regression	II-71
Basic Statistics for Rows	I-340
Basic Statistics	I-324
Bayesian Regression	II-99

Binary Logit with Interactions	III-33
Binary Logit with Multiple Predictors	III-27
Binary Logit with One Predictor	III-24
Binary Profiles	III-388
Bonferroni and Dunn-Sidak adjustments	I-552
Box-and-Whisker Plots	IV-166
Box-Behnken Design	I-380
Box-Cox Model	I-143
Box-Hunter Fractional Factorial Design	I-373
By-Choice Data Format	III-69
c Chart	IV-191
Calculating Percentiles Using Inverse Cumulative Distribution Function	IV-93
Calculating Probability Mass Function and Cumulative Distribution Function for Discrete Distributions	IV-90
Canonical Correlation Analysis	II-246
Canonical Correlations: Using Text Output	I-33
Canonical Correlations—Simple Model	IV-305
Casewise Pattern Table	III-142
Categorical Variables and Clustered Data	II-449
Central Composite Response Surface Design	I-384
Chi-Square Model for Signal Detection	IV-340

Choice Data	I-136
Circle Model	IV-11
Classical Test Analysis	IV-498
Classification Tree	I-55
Clustered Data in Mixed Regression	II-442
Cochran's Test of Linear Trend	I-273
Comparing Correlation Estimation Methods	III-168
Computation of p-value Using 1-CF Function	IV-94
Conditional Logistic Regression.	III-56
Confidence Curves and Regions.	III-287
Confidence Interval for Non-Centrality Parameter in One-Way Balanced Fixed Effect ANOVA.	IV-95
Confidence Intervals for Mean and Median.	I-28
Confidence Intervals for One-Way Table Percentages	I-250
Confidence Intervals for Smoothers.	IV-368
Confidence Intervals.	II-414
Contingency Table Analysis.	IV-312
Contouring the Loss Function	III-296
Contrasts	I-435
Correlation Estimation.	III-154

Correspondence Analysis (Simple)	I-207
Covariance Alternatives to Repeated Measures	II-234
Cox Regression	IV-462
Cross-Correlation Plot	IV-550
Crossover and Changeover Designs	II-222
Cross-Validation	I-444
Cross-Validation	III-371
Cumulative Histogram	IV-164
Cusum Charts	IV-201
Deciles of Risk and Model Diagnostics	III-39
Density Clustering Examples	I-112
Differencing	IV-552
Discrete Choice Models	III-60
Discriminant Analysis Using Automatic Backward Stepping	I-420
Discriminant Analysis Using Automatic Forward Stepping	I-413
Discriminant Analysis Using Complete Estimation	I-409
Discriminant Analysis Using Interactive Stepping	I-427
Discriminant Analysis	II-238
Employment Discrimination	I-147
Equality of Proportions	IV-63

Estimation: ML and REML	II-369
EWMA Chart	IV-204
Exploring with Residuals	II-334
Factor Analysis Using a Covariance Matrix.	I-482
Factor Analysis Using a Rectangular File	I-485
Fine Tuning	II-382
Fisher's Exact Test.	I-271
Fitting a Second Order Response Surface	IV-245
Fitting Binomial Distribution	I-504
Fitting Discrete Uniform Distribution	I-505
Fitting Exponential Distribution.	I-507
Fitting Gumbel Distribution	I-508
Fitting Multiple Distributions	I-513
Fitting Normal Distribution	I-510
Fitting Weibull Distribution	I-511
Fixing Parameters and Evaluating Fit	III-290
Flexible Beta Linkage Method for Hierarchical Clustering	I-115
Fourier Modeling of Temperature	IV-575
Fractional Factorial Design	I-372
Fractional Factorial Designs.	II-213

Frequency Input	I-256
Friedman Test for the Case with Ties	III-348
Friedman Test	III-347
From VC to MIXED	II-357
Full Factorial Designs	I-371
Functions of Parameters	III-293
Gamma Model for Signal Detection	IV-344
Geometric Mean	I-326
Getting Acquainted with the Output Layout	II-311
Guttman Loss Function.	III-198
Hadi Robust Outlier Detection	I-192
Harmonic Mean.	I-327
Heteroskedasticity-Consistent Standard Errors	IV-587
Hierarchical Clustering with Leaf Option	I-118
Hierarchical Clustering: Clustering Cases	I-105
Hierarchical Clustering: Clustering Variables and Cases	I-109
Hierarchical Clustering: Clustering Variables	I-108
Hierarchical Clustering: Distance Matrix Input	I-111
Histogram.	IV-163
Hotelling's T-Square	II-237

Hypothesis testing	II-372
Hypothesis Testing	III-77
Incomplete Block Designs.	II-212
Independent Samples t-Test	IV-72
Individual Differences Multidimensional Scaling.	III-200
Interactive Stepwise Regression	II-75
Internal Model	IV-12
Iterated Principal Axis	I-476
Iteratively Reweighted Least-Squares for Logistic Models	III-299
Kinetic Models.	III-313
K-Means Clustering	I-96
Kriging (Ordinary).	IV-411
Kruskal Method	III-195
Kruskal-Wallis Test	III-340
Latin Square Designs	II-220
Latin Squares	I-375
Least-Squares Regression	I-23
Life Tables; The Kaplan-Meier Estimator.	IV-449
Logistic Model (One Parameter)	IV-500
Logistic Model (Two Parameter)	IV-503

Logistic Model for Signal Detection	IV-335
Loglinear Modeling of a Four-Way Table	III-105
Longitudinal Data in Mixed Regression	II-457
LOWESS Smoothing	IV-558
Mann-Kendall test	IV-572
Mann-Whitney Test	III-342
Mantel-Haenszel Test	I-293
Maximum Likelihood Estimation	III-298
Maximum Likelihood	I-473
McNemar's Test of Symmetry	I-277
Minimizing an Analytic Function	III-315
Missing Category Codes	I-257
Missing Cells Designs (the Means Model)	II-224
Missing Data	II-340
Missing Data: EM Estimation	I-186
Missing Data: Pairwise Deletion	I-185
Missing Value Imputation	III-176
Missing Values: Preliminary Examinations	III-137
Mixture Design with Constraints	I-382
Mixture Design	I-381

Mixture Models	II-247
Moving Average Chart	IV-203
Moving Averages	IV-555
Multinomial Logit	III-50
Multiple Categories	III-390
Multiple Correspondence Analysis	I-214
Multiple Linear Regression	II-67
Multiple Response Optimization using Desirability Analysis.	IV-250
Multiplicative Seasonal Factor	IV-560
Multiplicative Seasonality with a Linear Trend	IV-561
Multivariate Layout for Longitudinal Data	II-473
Multivariate Nested Design	III-253
Multivariate Normality Assessment of Anthropometric Measurements	III-219
Multivariate Normality Assessment of Perspiration Measurements	III-218
Multivariate Regression by PLS Technique.	III-368
Multiway Tables	I-279
Negative Exponential Model for Signal Detection	IV-336
Nested Designs	II-215
Nested Effects	II-320
Nested Random Effects	II-417

Nesting in Design Structure	II-402
Nesting in treatment structure	II-399
Nesting versus Crossing	II-408
Nonlinear Model with Three Parameters	III-284
Nonmetric Unfolding	III-203
Nonparametric Model for Signal Detection	IV-333
Nonparametric: One Sample Kolmogorov-Smirnov Test Statistic.	I-36
Normal Distribution Model for Signal Detection	IV-328
Normality Assessment Using Shapiro-Wilk and Anderson-Darling Test	I-341
np Chart.	IV-183
N-tiles and P-tiles.	I-338
OC Curve for Binomial Distribution	IV-199
OC Curve for Variances	IV-198
OC Curve	IV-197
Odds Ratios.	I-269
One-Sample Kolmogorov-Smirnov Test for Non-Central Chi-square Distribution	III-352
One-Sample Kolmogorov-Smirnov Test for Normal Distribution.	III-350
One-Sample t-Test	I-547
One-Sample z-Test	I-544
One-Way ANOVA and Sample Size Estimation.	IV-77

One-Way ANOVA	II-122
One-Way ANOVA	II-203
One-Way MANOVA	III-246
One-Way Repeated Measures	II-155
One-Way Tables	I-248
Optimal Designs: Coordinate Exchange.	I-386
Optimizing Response using Canonical Analysis	IV-247
Optimum Choice of Number of Factors	III-375
Outliers in X-space and Y-space	IV-284
Outliers in X-space	IV-283
Outliers in Y-space	IV-280
p Chart	IV-189
Paired t-Test	I-548
Paired t-Test	IV-67
Pairwise comparisons	II-145
Pareto Charts.	IV-165
Partial Autocorrelation Plot	IV-549
Partial Correlations	II-248
Partial Set Correlation Model	IV-308
Path Analysis and Standard Errors	III-442

Path Analysis Basics	III-414
Path Analysis Using Rectangular Input	III-434
Path Analysis with a Restart File	III-419
PCA with Beta Distribution	IV-215
PCA With Box-Cox Transformation	IV-213
PCA with Normal Distribution	IV-212
PDL with Instrumental Variables	IV-596
PDL without Instrumental Variables	IV-595
Pearson Correlations	I-179
Percentages	I-258
Piecewise Regression.	III-311
Plackett-Burman Design	I-379
Point Statistics	IV-418
Poisson Model for Signal Detection	IV-342
Poisson Test	I-551
Polynomial Regression and Smoothing	IV-370
POSAC: Proportion of Profile Pairs Correctly Represented	I-34
Post hoc tests	II-379
Power Scaling Ratio Data	III-208
Prediction of New Observations	II-95

Principal Components Analysis (Within Groups)	II-242
Principal Components	I-469
Probabilities Associated with Correlations	I-188
Probit Analysis (Simple Model)	IV-104
Probit Analysis with Interactions	IV-106
Procrustes Rotation	IV-14
Quade Test for Cases with Ties	III-349
Quade Test for Multiple Comparisons.	III-349
Quadratic Model.	I-438
Quantiles.	III-45
R Chart.	IV-180
Randomized Block Designs	II-211
Regression Charts	IV-207
Regression Imputation.	III-181
Regression Tree with Box Plots	I-57
Regression Tree with Dit Plots	I-59
Regression using SSCP, Covariance or Correlation matrices	II-89
Regression with Ecological or Grouped Data	II-86
Regression without the Constant	II-87
Regression	III-306

Repeated Measures Analysis in the Presence of Subject-Specific Covariates	III-255
Repeated Measures Analysis of Covariance	II-170
Repeated Measures ANOVA for One Grouping Factor and One Within Factor with Ordered Levels.	II-160
Repeated Measures ANOVA for Two Grouping Factors and One Within Factor	II-163
Repeated Measures ANOVA for Two Trial Factors	II-166
Repeated Measures Experiment with Covariates.	II-366
Residuals and Diagnostics for Simple Linear Regression	II-63
Ridge Analysis	IV-249
Ridge Regression Analysis	II-97
Robust Discriminant Analysis	I-449
Robust Estimation (Measures of Location)	III-301
Rotation.	I-478
Run Chart.	IV-167
s chart.	IV-178
S2 and S3 Coefficients	I-196
Sampling Distribution of Double Exponential (Laplace) Median	IV-225
Saving Basic Statistics: Multiple Statistics and Grouping Variables	I-328
Saving Basic Statistics: One Statistic and One Grouping Variable	I-327
Scalogram Analysis—A Perfect Fit	III-386

Screening Effects	III-114
Seasonal Trend tests	IV-573
Seemingly Unrelated Regression Equations.	II-91
Separate Variance Hypothesis Tests.	II-151
Sign and Wilcoxon Tests for Multiple Variables	III-346
Sign Test.	III-343
Simple Correspondence Analysis using Raw Data	I-212
Simple Linear Regression	II-55
Simulation of Assembly System.	IV-226
Simulation	IV-417
Single-Degree-of-Freedom Designs	II-148
Smart Correspondence Analysis with Row-by-Column Data.	I-210
Smoothing (4253H Filter)	IV-557
Smoothing Binary Data in Three Dimensions.	IV-380
Smoothing: Saving and Plotting Results	IV-367
Spearman Correlations.	I-195
Spearman Rank Correlation	I-27
Split Plot Design	II-323
Split Plot Designs	II-217
Stem-and-Leaf Plot for Rows	I-342

Stern-and-Leaf Plot	I-333
Stepwise Regression	III-70
Stepwise Regression	IV-468
Stratified Cox Regression	IV-464
Stratified Kaplan-Meier Estimation	IV-455
Structural Zeros	III-117
Structured Covariance Matrix for Random Errors	II-362
Tables with Ordered Categories	I-275
Tables without Analyses	III-121
Tackling different data format in Logistic Regression	III-81
Taguchi Design	I-377
Test for Equality of Several Variances	I-558
Test for Equality of Two Correlation Coefficients	I-562
Test for Equality of Two Proportions	I-564
Test for Equality of Two Variances	I-557
Test for Single Proportion	I-564
Test for Single Variance	I-556
Test for Specific Correlation Coefficient	I-560
Test for Zero Correlation Coefficient	I-559
Testing Nonzero Null Hypotheses	II-85

Testing whether a Single Coefficient Equals Zero	II-81
Testing whether Multiple Coefficients Equal Zero	II-83
Tetrachoric Correlation	I-198
The Nelson-Aalen Estimator	IV-451
The Weibull Model for Fully Parametric Analysis	IV-472
Time Series Plot	IV-547
Transformations	I-182
Transformations	II-60
Treatment or design?	II-406
TSLS without lag and with hypothesis testing	IV-593
TSQ Chart	IV-209
Turnbull Estimation: K-M for Interval-Censored Data	IV-459
Two-Sample t-Test	I-549
Two-Sample z-Test	I-545
Two-Stage Instrumental Variables	IV-592
Two-Stage Least Squares	IV-590
Two-Way MANOVA	III-248
Two-Way ANOVA	II-132
Two-way ANOVA	IV-80
Two-Way Table Measures (Long Results)	I-263

Two-Way Table Measures	I-261
Two-Way Tables	I-253
u Chart	IV-195
Unbalanced ANOVA	II-146
Unbalanced Data: Different Types of ANOVA	II-328
Univariate Regression by PLS Technique	III-365
Unordered Data	I-198
Unusual Distances	IV-424
Usefulness of Jackknife estimate	I-30
Using Covariates	II-326
Validity indices RMSSTD, Pseudo F, and Pseudo T-square with cities	I-116
Variance Chart	IV-176
Vector Model	IV-9
Wald-Wolfowitz Runs Test	III-354
Weighting Means	II-234
Wilcoxon Test	III-345
Within-Group Testing	III-257
Word Frequency	I-140
X-bar Chart	IV-168
X-MR Chart (Sigma Estimation with Median).	IV-206

Linear Models

This chapter in this manual normally has a very technical feel to it. In the past, however, Regression, ANOVA, and General Linear Models are grouped together. There are two reasons for doing this. First, a single set of statistical procedures treat regression and analysis of variance as related, rather than as distinct. These are based on the same underlying mathematics. When you study what these procedures do, therefore, it is helpful to understand the model and learn the basic statistical methodology underlying each method. Second, although SYSTAT has three commands (REGRESS, ANOVA, and GLM) for these subjects, it is a not-so-well-guarded secret that all these lead to the same program, originally called MOLH (for Multivariate General Linear Hypothesis). Having been convinced it is very rare that SYSTAT can use tools designed for one approach (for example, dummy variables in ANOVA) in another (such as computing within-group correlations by multivariate regression). This synergy is not usually available to packages that treat these models independently.

Simple Linear Models

Linear models are models based on lines. More generally, they are based on linear surfaces, such as lines, planes, and hyperplanes. Linear models are widely applied because lines and planes often appear to describe well the random, noisy, variation encountered in the real world. We will begin by estimating the equation for a simple line and then move to more complex linear models.

Linear Models

Each chapter in this manual normally has its own statistical background section. In this part, however, Regression, ANOVA, and General Linear Models are grouped together. There are two reasons for doing this. First, while some introductory textbooks treat regression and analysis of variance as distinct, statisticians know that they are based on the same underlying mathematical model. When you study what these procedures do, therefore, it is helpful to understand this model and learn the common terminology underlying each method. Second, although SYSTAT has three commands (REGRESS, ANOVA, and GLM) and menu settings, it is a not-so-well-guarded secret that all these lead to the same program, originally called MGLH (for Multivariate General Linear Hypothesis). Having them organized this way means that SYSTAT can use tools designed for one approach (for example, dummy variables in ANOVA) in another (such as computing within-group correlations in multivariate regression). This synergy is not usually available in packages that treat these models independently.

Simple Linear Models

Linear models are models based on *lines*. More generally, they are based on linear surfaces, such as lines, planes, and hyperplanes. Linear models are widely applied because lines and planes often appear to describe well the relations among variables measured in the real world. We will begin by examining the equation for a straight line, and then move to more complex linear models.

Equation for a Line

A linear model looks like this:

$$y = a + bx$$

This is the equation for a straight line that you learned in school. The quantities in this equation are:

- y a dependent variable
- x an independent variable

Variables are quantities that can vary (have different numerical values) in the same equation. The remaining quantities are called **parameters**. A parameter is a quantity that is constant in a particular equation, but that can be varied to produce other equations in the same general family. The parameters are:

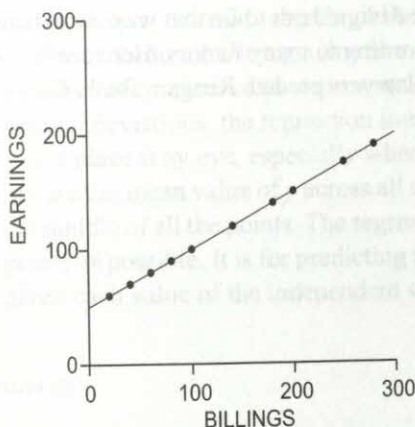
- a The value of y when x is 0. This is sometimes called a y -intercept (where a line intersects the y axis in a graph when x is 0).
- b The slope of the line, or the number of units y changes when x changes by one unit.

Let us look at an example. Here are some data showing the yearly earnings a partner should theoretically get in a certain large law firm, based on annual personal billings over quota (both in thousands of dollars):

<i>EARNINGS</i>	<i>BILLINGS</i>
60	20
70	40
80	60
90	80
100	100
120	140
140	180
150	200
175	250
190	280

We can plot these data with *EARNINGS* on the vertical axis (dependent variable) and *BILLINGS* on the horizontal (independent variable). Notice in the following figure that

all the points lie on a straight line.



What is the equation for this line? Look at the vertical axis value on the sloped line where the independent variable has a value of 0. Its value is 50. A lawyer is paid \$50,000 even when billings nothing. Thus, a is 50 in our equation. What is b ? Notice that the line rises by \$10,000 when billings change by \$20,000. The line rises half as fast as it runs. You can also look at the data and see that the earnings change by \$1 as billing changes by \$2. Thus, b is 0.5, or a half, in our equation.

Why bother with all these calculations? We could use the table to determine a lawyer's compensation, but the formula and the line graph allow us to determine wages *not* found in the table. For example, we now know that \$30,000 in billings would yield earnings of \$65,000:

$$EARNINGS = 50000 + 0.5 \times 30000 = 65000$$

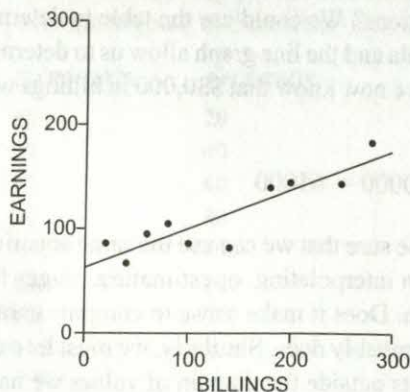
When we do this, however, we must be sure that we can use the same equation on these new values. We must be careful when interpolating, or estimating, wages for billings between the ones we have been given. Does it make sense to compute earnings for \$25,000 in billings, for example? It probably does. Similarly, we must be careful when extrapolating, or estimating from units outside the domain of values we have been given. What about negative billings, for example? Would we want to pay an embezzler? Be careful. Equations and graphs usually are meaningful only within or close to the range of y values and domain of x values in the data.

Regression

Data are seldom this clean unless we design them to be that way. Law firms typically fine tune their partners' earnings according to many factors. Here are the real billings and earnings for our law firm (these lawyers predate Reagan, Bush, Clinton, and Gates):

EARNINGS	BILLINGS
86	20
67	40
95	60
105	80
86	100
82	140
140	180
145	200
144	250
184	280

Our techniques for computing a linear equation won't work with these data. Look at the following graph. There is no way to draw a straight line through all the data.



Given the irregularities in our data, the line drawn in the figure is a compromise. How do we find a best fitting line? If we are interested in predicting earnings from the billing data values rather well, a reasonable method would be to place a line through the points

so that the vertical deviations between the points and the line (errors in predicting earnings) are as small as possible. In other words, these deviations (absolute discrepancies, or **residuals**) should be small, on the average, for a good-fitting line.

The procedure of fitting a line or curve to data such that residuals on the dependent variable are minimized in some way is called **regression**. Because we are minimizing vertical deviations, the regression line often appears to be more horizontal than we might place it by eye, especially when the points are fairly scattered. It “regresses” toward the mean value of y across all the values of x , namely, a horizontal line through the middle of all the points. The regression line is not intended to pass through as many points as possible. It is for predicting the dependent variable as accurately as possible, given each value of the independent variable.

Least Squares

There are several ways to draw the line so that, on the average, the deviations are small. We could minimize the mean, the median, or some other measure of the typical behavior of the absolute values of the residuals. Or we can minimize the sum (or mean) of the squared residuals, which yields almost the same line in most cases. Using squared instead of absolute residuals gives more influence to points whose y value is farther from the average of all y values. This is not always desirable, but it makes the mathematics simpler. This method is called **ordinary least squares**.

By specifying *EARNINGS* as the dependent variable and *BILLINGS* as the independent variable in a **MODEL** statement, we can compute the ordinary least-squares regression y -intercept as \$62,800 and the slope as 0.375. These values do not predict any single lawyer’s earnings exactly. They describe the whole firm well, in the sense that, on the average, the line predicts a given earnings value fairly closely from a given billings value.

Estimation and Inference

We often want to do more with such data than draw a line on a picture. In order to generalize, formulate a policy, or test a hypothesis, we need to make an **inference**. Making an inference implies that we think a model describes a more general population from which our data have been randomly sampled. In the present example, this population is all possible lawyers who might work for this firm. To make an inference about compensation, we need to construct a linear model for our population that includes a parameter for random error. In addition, we need to change our notation

to avoid confusion later. We are going to use Greek letters to denote parameters and italic Roman letters for variables. The error parameter is usually called ε .

$$y = \alpha + \beta x + \varepsilon$$

Notice that ε is a **random variable**. It varies like any other variable (for example, x), but it varies randomly, like the tossing of a coin. Since ε is random, our model forces y to be random as well because adding fixed values (α and βx) to a random variable produces another random variable. In ordinary language, we are saying with our model that earnings are only partly predictable from billings. They vary slightly according to many other factors, which we assume are random.

We do not know all of the factors governing the firm's compensation decisions, but we assume:

- All the salaries are derived from the same linear model.
- The error in predicting a particular salary from billings using the model is independent of (not in any way predictable from) the error in predicting other salaries.
- The errors in predicting all the salaries come from the same random distribution.

Our model for predicting in our population contains parameters, but unlike our perfect straight line example, we cannot compute these parameters directly from the data. The data we have are only a small sample from a much larger population, so we can only estimate the parameter values using some statistical method on our sample data. Those of you who have heard this story before may not be surprised that ordinary least squares is one reasonable method for estimating parameters when our three assumptions are appropriate. Without going into all the details, we can be reasonably assured that if our population assumptions are true and if we randomly sample some cases (that is, each case has an equal chance of being picked) from the population, the least-squares estimates of α and β will, on the average, be close to their values in the population.

So far, we have done what seems like a sleight of hand. We delved into some abstruse language and came up with the same least-squares values for the slope and intercept as before. There is something new, however. We have now added conditions that define our least-squares values as sample estimates of population values. We now regard our sample data as one instance of many possible samples. Our compensation model is like Plato's cave metaphor; we think it typifies how this law firm makes compensation decisions about any lawyer, not just the ones we sampled. Before, we were computing *descriptive statistics* about a sample. Now, we are computing *inferential statistics* about a population.

Standard Errors

There are several statistics relevant to the estimation of α and β . Perhaps most important is a measure of how variable we could expect our estimates to be if we continued to sample data from our population and used least squares to get our estimates. A statistic calculated by SYSTAT shows what we could expect this variation to be. It is called, appropriately, the **standard error of estimate**, or *Std Error* in the output. The standard error of the y -intercept, or regression constant, is in the first row of the coefficients: 10.440. The standard error of the billing coefficient or slope is 0.065. Look for these numbers in the following output:

Dependent Variable	EARNINGS
N	10
Multiple R	0.897
Squared Multiple R	0.804
Adjusted Squared Multiple R	0.779
Standard Error of Estimate	17.626

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-value
CONSTANT	62.838	10.440	0.000	.	6.019	0.000
BILLINGS	0.375	0.065	0.897	1.000	5.728	0.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	10191.109	1	10191.109	32.805	0.000
Residual	2485.291	8	310.661		

Hypothesis Testing

From these standard errors, we can construct hypothesis tests on these coefficients. Suppose a skeptic approached us and said, "Your estimates look as if something is going on here, but in this firm, salaries have nothing to do with billings. You just happened to pick a sample that gives the impression that billings matter. It was the luck of the draw that provided you with such a misleading picture. In reality, β is 0 in the population because billings play no role in determining earnings."

We can reply, "If salaries had nothing to do with billings but are really just a mean value plus random error for any billing level, then would it be likely for us to find a coefficient estimate for β at least this different from 0 in a sample of 10 lawyers?"

To represent these alternatives as a bet between us and the skeptic, we must agree on some critical level for deciding who will win the bet. If the likelihood of a sample

result at least this extreme occurring by chance is less than or equal to this critical level (say, five times out of a hundred), we win; otherwise, the skeptic wins.

This logic might seem odd at first because, in almost every case, our skeptic's *null hypothesis* would appear ridiculous, and our *alternative hypothesis* (that the skeptic is wrong) seems plausible. Two scenarios are relevant here, however. The first is the lawyer's. We are trying to make a case here. The only way we will prevail is if we convince our skeptical jury beyond a reasonable doubt. In statistical practice, that reasonable doubt level is relatively liberal: fewer than five times in a hundred. The second scenario is the scientist's. We are going to stake our reputation on our model. If someone sampled new data and failed to find nonzero coefficients, much less coefficients similar to ours, few would pay attention to us in the future.

To compute probabilities, we must count all possibilities or refer to a mathematical probability distribution that approximates these possibilities well. The most widely used approximation is the normal curve, which we reviewed briefly in Chapter 1 in Statistics I. For large samples, the regression coefficients will tend to be normally distributed under the assumptions we made above. To allow for smaller samples, however, we will add the following condition to our list of assumptions:

- The errors in predicting the salaries come from a normal distribution.

If we estimate the standard errors of the regression coefficients from the data instead of knowing them in advance, then we should use the *t* distribution instead of the normal. The two-tail value for the probability represents the area under the theoretical *t* probability curve corresponding to coefficient estimates whose absolute values are more extreme than the ones we obtained. For both parameters in the model of lawyers' earnings, these values (given as *p-value(2 tail)*) are less than 0.001, leading us to reject our null hypothesis at well below the 0.05 level.

At the bottom of our output, we get an analysis of variance table that tests the goodness of fit of our entire model. The null hypothesis corresponding to the *F-ratio* (32.805) and its associated *p-value* is that the billing variable coefficient is equal to 0. This test overwhelmingly rejects the null hypothesis that both α and β are 0.

Multiple Correlation

In the same output is a statistic called the **squared multiple correlation**. This is the proportion of the total variation in the dependent variable (*EARNINGS*) accounted for by the linear prediction using *BILLINGS*. The value here (0.804) tells us that approximately 80% of the variation in earnings can be accounted for by a linear prediction from billings.

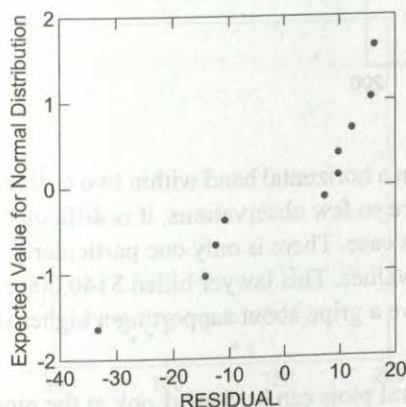
The rest of the variation, as far as this model is concerned, is random error. The square root of this statistic is called, not surprisingly, the **multiple correlation**. The adjusted squared multiple correlation (0.779) is what we would expect the squared multiple correlation to be if we used the model we just estimated on a new sample of 10 lawyers in the firm. It is smaller than the squared multiple correlation because the coefficients were optimized for this sample rather than for the new one.

Regression Diagnostics

We do not need to understand the mathematics of how a line is fitted in order to use regression. You can fit a line to any x - y data by the method of least-squares. The computer doesn't care where the numbers come from. To have a model and estimates that mean something, however, you should be sure the assumptions are reasonable and that the sample data appear to be sampled from a population that meets the assumptions.

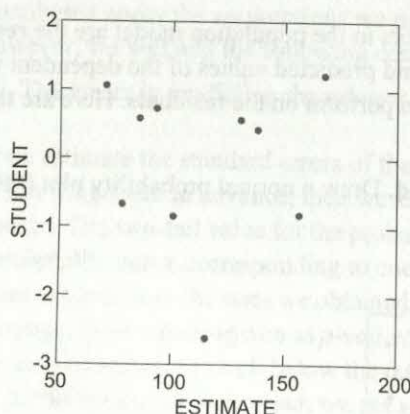
The sample analogues of the errors in the population model are the **residuals**—the differences between the observed and predicted values of the dependent variable. There are many diagnostics you can perform on the residuals. Here are the most important ones:

The errors are normally distributed. Draw a normal probability plot (PLOT) of the residuals.



The residuals should fall approximately on a diagonal straight line in this plot. When the sample size is small, as in our law example, the line may be quite jagged. It is difficult to tell by any method whether a small sample is from a normal population. You can also plot a histogram or stem-and-leaf diagram of the residuals to see if they are lumpy in the middle with thin, symmetric tails.

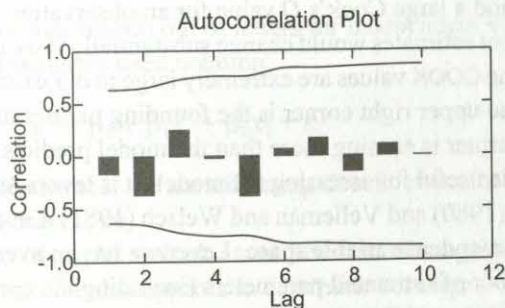
The errors have constant variance. Plot the residuals against the estimated values. The following plot shows studentized residuals (*STUDENT*) against estimated values (*ESTIMATE*). Studentized residuals are the true “external” kind discussed in Velleman and Welsch (1981). Use these statistics to identify outliers in the dependent variable space. Under normal regression assumptions, they have a t distribution with $(N - p - 1)$ degrees of freedom, where N is the total sample size and p is the number of predictors (including the constant). Large values (greater than 2 or 3 in absolute magnitude) indicate possible problems.



Our residuals should be arranged in a horizontal band within two or three units around 0 in this plot. Again, since there are so few observations, it is difficult to tell whether they violate this assumption in this case. There is only one particularly large residual, and it is toward the middle of the values. This lawyer billed \$140,000 and is earning only \$80,000. He or she might have a gripe about supporting a higher share of the firm's overhead.

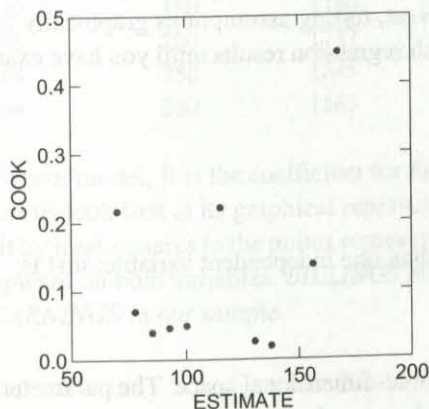
The errors are independent. Several plots can be done. Look at the plot of residuals against estimated values above. Make sure that the residuals are randomly scattered above and below the 0 horizontal and that they do not track in a snaky way across the

plot. If they look as if they were shot at the plot by a horizontally moving machine gun, then they are probably not independent of each other. You may also want to plot residuals against other variables, such as time, orientation, or other ways that might influence the variability of your dependent measure. ACF PLOT in SERIES measures whether the residuals are *serially correlated*. Here is an autocorrelation plot:



All the bars should be within the confidence bands if each residual is not predictable from the one preceding it, and the one preceding that, and the one preceding that, and so on.

All the members of the population are described by the same linear model. Plot Cook's distance (COOK) against the estimated values.



Cook's distance measures the influence of each sample observation on the coefficient estimates. Observations that are far from the average of all the independent variable values or that have large residuals tend to have a large Cook's distance value (say, greater than 2). Cook's D actually closely follows an F distribution, so aberrant values depend on the sample size. As a rule of thumb, under the normal regression assumptions, COOK can be compared to an F distribution with p and $N - p$ degrees of freedom. We don't want to find a large Cook's D value for an observation because it would mean that the coefficient estimates would change substantially if we deleted that observation. While none of the COOK values are extremely large in our example, could it be that the largest one in the upper right corner is the founding partner in the firm? Despite large billings, this partner is earning more than the model predicts.

Another diagnostic statistic useful for assessing the model fit is leverage, discussed in Belsley, Kuh, and Welsch (1980) and Velleman and Welsch (1981). Leverage helps to identify outliers in the independent variable space. Leverage has an average value of p/N , where p is the number of estimated parameters (including the constant) and N is the number of cases. What is a high value of leverage? In practice, it is useful to examine the values in a stem-and-leaf plot and identify those that stand apart from the rest of the sample. However, various rules of thumb have been suggested. For example, values of leverage less than 0.2 appear to be safe; between 0.2 and 0.5, risky; and above 0.5, to be avoided. Another says that if $p > 6$ and $(N - p) > 12$, use $(3p)/N$ as a cutoff. SYSTAT uses an F approximation to determine this value for warnings (Belsley, Kuh, and Welsch, 1980).

In conclusion, keep in mind that all our diagnostic tests are themselves a form of inference. We can assess theoretical errors only through the dark mirror of our observed residuals. Despite this caveat, testing assumptions graphically is critically important. You should never publish regression results until you have examined these plots.

Multiple Regression

A multiple linear model has more than one independent variable; that is:

$$y = a + bx + cz$$

This is the equation for a plane in three-dimensional space. The parameter a is still an intercept term. It is the value of y when x and z are 0. The parameters b and c are still

slopes. One gives the slope of the plane along the x dimension; the other, along the z dimension.

The statistical model has the same form:

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

Before we run out of letters for independent variables, let us switch to a more frequently used notation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

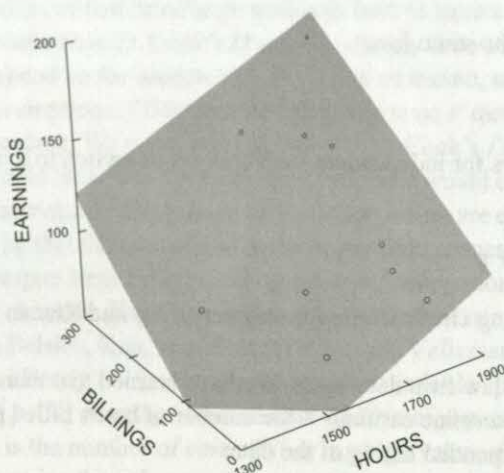
Notice that we are still using Greek letters for unobservables and Roman letters for observables.

Now, let us look at our law firm data again. We have learned that there is another variable that appears to determine earnings—the number of hours billed per year by each lawyer. Here is an expanded listing of the data:

<i>EARNINGS</i>	<i>BILLINGS</i>	<i>HOURS</i>
86	20	1771
67	40	1556
95	60	1749
105	80	1754
86	100	1594
82	140	1400
140	180	1780
145	200	1737
144	250	1645
184	280	1863

For our model, β is the coefficient for *BILLINGS*, and β is the coefficient for *HOURS*. Let us look first at its graphical representation. The following figure shows the plane fit by least-squares to the points representing each lawyer. Notice how the plane slopes upward on both variables. *BILLINGS* and *HOURS* both contribute positively to *EARNINGS* in our sample.

Fitted Model Plot



Fitting this model involves no more work than fitting the simple regression model. We specify one dependent and two independent variables and estimate the model as before. Here is the result:

Dependent Variable	EARNINGS
N	10
Multiple R	0.998
Squared Multiple R	0.996
Adjusted Squared Multiple R	0.995
Standard Error of Estimate	2.678

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t	p-value
CONSTANT	-139.925	11.116	0.000	.	-12.588	0.000
BILLINGS	0.333	0.010	0.797	0.951	32.690	0.000
HOURS	0.124	0.007	0.449	0.951	18.429	0.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	12626.210	2	6313.105	880.493	0.000
Residual	50.190	7	7.170		

This time, we have one more row in our regression table for *HOURS*. Notice that its coefficient (0.124) is smaller than that for *BILLINGS* (0.333). This is due partly to the different scales of the variables. *HOURS* are measured in larger numbers than *BILLINGS*. If we wish to compare the influence of each, independent of scales, we should look at the standardized coefficients. Here, we still see that *BILLINGS* (0.797) play a greater role in predicting *EARNINGS* than do *HOURS* (0.449). Notice also that both coefficients are highly significant and that our overall model is highly significant, as shown in the analysis of variance table.

Variable Selection

In applications, you may not know which subset of predictor variables in a larger set constitutes a “good” model. Strategies for identifying a good subset are many and varied: forward selection, backward elimination, stepwise (either a forward or backward type), and all subsets. Forward selection begins with the “best” predictor, adds the next “best” and continues entering variables to improve the fit. Backward selection begins with all candidate predictors in an equation and removes the least useful one at a time as long as the fit is not substantially “worsened.” Stepwise begins as either forward or backward, but allows “poor” predictors to be removed from the candidate model or “good” predictors to re-enter the model at any step. Finally, all subsets methods compute all possible subsets of predictors for each model of a given size (number of predictors) and choose the “best” one.

Bias and variance tradeoff. Submodel selection is a tradeoff between bias and variance. By decreasing the number of parameters in the model, its predictive capability is enhanced. This is because the variance of the parameter estimates decreases. On the other side, bias may increase because the “true model” may have a higher dimension. So we’d like to balance smaller variance against increased bias. There are two aspects to variable selection: selecting the dimensionality of the submodel (how many variables to include) and evaluating the model selected. After you determine the dimension, there may be several alternative subsets that perform equally well. Then, knowledge of the subject matter, how accurately individual variables are measured, and what a variable “communicates” may guide the selection of the model to report.

A strategy. If you are in an exploratory phase of research, you might try this version of backwards stepping. First, fit a model using all candidate predictors. Then identify the least “useful” variable, remove it from the model list, and fit a smaller model. Evaluate

your results and select another variable to remove. Continue removing variables. For a given size model, you may want to remove alternative variables (that is, first remove variable A , evaluate results, replace A and remove B , etc.).

Entry and removal criteria. Decisions about which variable to enter or remove should be based on statistics and diagnostics in the output, especially graphical displays of these values, and your knowledge of the problem at hand.

You can specify your own alpha-to-enter and alpha-to-remove values (do not make alpha-to-remove less than alpha-to-enter), or you can cycle variables in and out of the equation (stepping automatically stops if this happens). The default values for these options are Enter = 0.15 and Remove = 0.15. These values are appropriate for predictor variables that are relatively independent. If your predictor variables are highly correlated, you should consider lowering the Enter and Remove values well below 0.05.

When there are high correlations among the independent variables, the estimates of the regression coefficients can become unstable. Tolerance is a measure of this condition. It is $(1 - R^2)$; that is, one minus the squared multiple correlation between a predictor and the other predictors included in the model. (Note that the dependent variable is not used.) By setting a minimum tolerance value, variables highly correlated with others already in the model are not allowed to enter.

As a rough guideline, consider models that include only variables that have absolute t values well above 2.0 and “tolerance” values greater than 0.1. (We use quotation marks here because t and other statistics do not have their usual distributions when you are selecting subset models.)

Evaluation criteria. There is no one test to identify the dimensionality of the best submodel. Research by Leo Breiman emphasizes the usefulness of cross-validation techniques involving 80% random subsamples. Sample 80% of your file, fit a model, use the resulting coefficients on the remaining 20% to obtain predicted values, and then compute R^2 for this smaller sample. In over-fitting situations, the discrepancy between the R^2 for the 80% sample and the 20% sample can be dramatic.

A warning. If you do not have extensive knowledge of your variables and expect this strategy to help you to find a “true” model, you can get into a lot of trouble. Automatic stepwise regression programs cannot do your work for you. You must be able to examine graphics and make intelligent choices based on theory and prior knowledge; otherwise, you will be arriving at nonsense.

Moreover, if you are thinking of testing hypotheses after automatically fitting a subset model, don’t bother. Stepwise regression programs are the most notorious source of “pseudo” p -values in the field of automated data analysis. Statisticians seem

to be the only ones who know these are not “real” *p-values*. The automatic stepwise option is provided to select a subset model for prediction purposes. It should never be used without cross-validation.

If you still want some sort of confidence estimate on your subset model, you might look at tables in Wilkinson (1979), Rencher and Pun (1980), and Wilkinson and Dallal (1982). These tables provide null hypothesis R^2 values for selected subsets given the number of candidate predictors and final subset size. If you don't know this literature already, you will be surprised at how large multiple correlations from stepwise regressions on random data can be. For a general summary of these and other problems, see Hocking (1983). For more specific discussions of variable selection problems, see the previous references and Flack and Chang (1987), Freedman (1983), and Lovell (1983). Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need automated stepwise regression to solve a particular problem, it is almost certain that you do not. Professional statisticians rarely use automated stepwise regression because it does not necessarily find the “best” fitting model, the “real” model, or alternative “plausible” models. Furthermore, the order in which variables enter or leave a stepwise program is usually of no theoretical significance. You are always better off thinking about why a model could generate your data and then testing that model, AIC and Schwarz's BIC. Model selection criteria like likelihood and multiple- R^2 are biased towards models with more parameters, leading to over-fitted and less precise models.

Akaike (1973, 1974), proposed the Akaike Information Criterion (AIC) as a model selection criterion as follows:

$AIC = -2\text{Log-likelihood} + 2k$, where k is the number of parameters estimated.

Model selection using AIC is based on the principle of parsimony. AIC penalizes the likelihood with respect to the number of parameters estimated. The AIC value of a model can be interpreted as an estimate of the relative discrepancy between the model and the unknown true model which generated the data. The idea of model selection using AIC is to select a model with a low AIC value. Model selection using AIC is asymptotically equivalent to model selection by cross-validation. Proper care should be taken for model selection using AIC, and the AIC values should be used to compare models based on the same data and the same response. AIC may perform poorly if the number of parameters estimated is more relative to the number of observations.

Hurvich and Tsai (1989), provided small sample Akaike information criterion called, AIC (corrected) as follows;

$AIC(\text{corrected}) = -2\text{Log-likelihood} + 2k + 2k(k+1)/(n-k-1)$, where n is the number of observations. AIC (corrected) is applicable only for linear models with the underlying distribution being Gaussian.

Schwarz (1978) provided a Bayesian Information Criterion (BIC) for model selection.

Schwarz's $BIC = -2 * \text{Log-likelihood} + k * \log(n)$. Schwarz's BIC value of a model can also be interpreted as an estimate of relative discrepancy between the model and the unknown true model which generated the data. The idea is to select a model with a low Schwarz's BIC value.

Burnham and Anderson (2003) is a good source of material on information criteria and model selection.

In linear regression, ANOVA, GLM, and MANOVA the Log-likelihood is obtained under the assumption of normality.

In SYSTAT AIC, AIC (corrected) and Schwarz's BIC values are provided in Linear Regression (Least-Squares), ANOVA, GLM, Logit Regression, Probit Regression, Survival Analysis and MANOVA features. In Linear regression, ANOVA, GLM, and MANOVA the log-likelihood is obtained under the assumption of normality.

Using an SSCP, a Covariance, or a Correlation Matrix as Input

Normally for a regression analysis, you use a cases-by-variables data file. You can, however, use a covariance or correlation matrix saved (from Correlations) as input. If you use a matrix as input, specify the sample size that generated the matrix where the number you type is an integer greater than two.

You can enter an SSCP, a covariance, or a correlation matrix by typing it into the Data Editor Worksheet, by using BASIC, or by saving it in a SYSTAT file. Be sure to include the dependent as well as independent variables.

SYSTAT needs the sample size to calculate degrees of freedom, so you need to enter the original sample size. Least-Squares determines the type of matrix (SSCP, covariance, etc.) and adjusts appropriately. With a correlation matrix, the raw and standardized coefficients are the same. Therefore, the Include constant option is disabled when using SSCP, covariance, or correlation matrices. Because these matrices are centered, the constant term has already been removed.

The following two analyses of the same data file produce identical results (except that you don't get residuals with the second). In the first, we use the usual cases-by-variables data file. In the second, we use the CORR command to save a covariance matrix and then analyze that matrix file with the REGRESS command.

Here are the usual instructions for a regression analysis:

```
REGRESS
USE FILENAME
MODEL Y = CONSTANT + X(1) + X(2) + X(3)
ESTIMATE
```

Here, we compute a covariance matrix and use it in the regression analysis:

```
CORR
USE FILENAME1
SAVE filename2
COVARIANCE X(1) X(2) X(3) Y

REGRESS
USE FILENAME2
MODEL Y = X(1) + X(2) + X(3) / N=40
ESTIMATE
```

The triangular matrix input facility is useful for “meta-analysis” of published data and missing-value computations. There are a few warnings, however. First, if you input correlation matrices from textbooks or articles, you may not get the same regression coefficients as those printed in the source. Because of round-off error, printed and raw data can lead to different results. Second, if you use pairwise deletion with CORR, the degrees of freedom for hypotheses will not be appropriate. You may not even be able to estimate the regression coefficients because of singularities.

In general, when an incomplete data procedure is used to estimate the correlation matrix, the estimate of regression coefficients and hypothesis tests produced from it are optimistic. You can correct for this by specifying a sample size smaller than the number of actual observations (preferably, set it equal to the smallest number of cases used for any pair of variables), but this is a crude guess that you could refine only by doing Monte Carlo simulations. There is no simple solution. Beware, especially, of multivariate regressions (or MANOVA, etc.) with missing data on the dependent variables. You can usually compute coefficients, but results from hypothesis tests are particularly suspect.

Analysis of Variance

Often, you will want to examine the influence of categorical variables (such as gender, species, country, and experimental group) on continuous variables. The model equations for this case, called **analysis of variance**, are equivalent to those used in

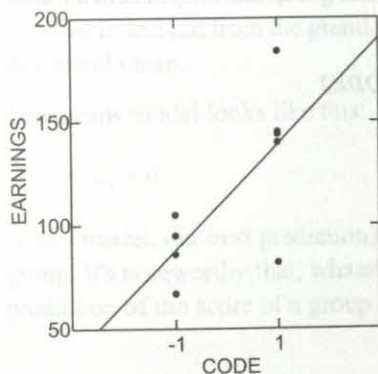
linear regression. However, in the latter, you have to figure out a numerical coding for categories so that you can use the codes in an equation as the independent variable(s).

Effects Coding

The following data file, *EARNBILL*, shows the breakdown of lawyers sampled by sex. Because *SEX* is a categorical variable (numerical values assigned to *MALE* or *FEMALE* are arbitrary), a code variable with the values 1 or -1 is used. It doesn't matter which group is assigned -1, as long as the other is assigned 1.

<i>EARNINGS</i>	<i>SEX</i>	<i>CODE</i>
86	female	-1
67	female	-1
95	female	-1
105	female	-1
86	female	-1
82	male	1
140	male	1
145	male	1
144	male	1
184	male	1

There is nothing wrong with plotting earnings against the code variable, as long as you realize that the slope of the line is arbitrary because it depends on how you assign your codes. By changing the values of the code variable, you can change the slope. Here is a plot with the least-squares regression line superimposed.



Let us do a regression on the data using these codes. Here are the coefficients as computed by ANOVA:

Variable Coefficients	
Constant	113.400
Code	25.600

Notice that *Constant* (113.4) is the mean of all the data. It is also the regression intercept because the codes are symmetrical about 0. The coefficient for *Code* (25.6) is the slope of the line. It is also one half the difference between the means of the groups. This is because the codes are exactly two units apart. This slope is often called an **effect** in the analysis of variance because it represents the amount that the categorical variable *SEX* affects *BILLINGS*. In other words, the effect of *SEX* can be represented by the amount that the mean for males differs from the overall mean.

Means Coding

The effects coding model is useful because the parameters (constant and slope) can be interpreted as an overall level and as the effect(s) of treatment, respectively. Another



106,2010
13999

model, however, that yields the means of the groups directly is called the means model. Here are the codes for this model:

<i>EARNINGS</i>	<i>SEX</i>	<i>CODE1</i>	<i>CODE2</i>
86	female	1	0
67	female	1	0
95	female	1	0
105	female	1	0
86	female	1	0
82	male	0	1
140	male	0	1
145	male	0	1
144	male	0	1
184	male	0	1

Notice that *CODE1* is nonzero for all females, and *CODE2* is nonzero for all males. To estimate a regression model with these codes, you must leave out the constant. With only two groups, only two distinct pieces of information are needed to distinguish them. Here are the coefficients for these codes in a model without a constant:

Variable	Coefficient
Code1	87.800
Code2	139.000

Notice that the coefficients are now the means of the groups.

Models

Let us look at the algebraic models for each of these codings. Recall that the regression model looks like this:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

For the effects model, it is convenient to modify this notation as follows:

$$y_j = \mu + \alpha_j + \varepsilon$$

When x (the code variable) is -1 , α_j is equivalent to α_1 ; when x is 1 , α_j is equivalent to α_2 . This shorthand will help you later when dealing with models with many categories. For this model, the μ parameter stands for the grand (overall) mean, and the α

parameter stands for the effect. In this model, our best prediction of the score of a group member is derived from the grand mean plus or minus the deviation of that group from this grand mean.

The means model looks like this:

$$y_j = \mu_j + \varepsilon$$

In this model, our best prediction of the score of a group member is the mean of that group. It's noteworthy that, whatever be the coding (means, effect, or dummy) the prediction of the score of a group member remains the same.

Hypotheses

As with regression, we are usually interested in testing hypotheses concerning the parameters of the model. Here are the hypotheses for the two models:

$$H_0: \alpha_1 = \alpha_2 = 0 \text{ (effects model)}$$

$$H_0: \mu_1 = \mu_2 \text{ (means model)}$$

The tests of this hypothesis compare variation between the means to variation within each group, which is mathematically equivalent to testing the significance of coefficients in the regression model. In our example, the *F-ratio* in the analysis of variance table tells you that the coefficient for *SEX* is significant at $p = 0.019$, which is less than the conventional 0.05 value. Thus, on the basis of this sample and the validity of our usual regression assumptions, you can conclude that women earn significantly less than men in this firm.

Dependent Variable	EARNINGS
N	10
Multiple R	0.719
Squared Multiple R	0.517

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
SEX\$	6553.600	1	6553.600	8.563	0.019
Error	6122.800	8	765.350		

The nice thing about realizing that ANOVA is specially-coded regression is that the usual assumptions and diagnostics are appropriate in this context. You can plot residuals against estimated values, for example, to check for homogeneity of variance.

Multigroup ANOVA

When there are more groups, the coding of categories becomes more complex. For the effects model, there are one fewer coding variables than number of categories. For two categories, you need only one coding variable; for three categories, you need two coding variables:

Category Code

1	1	0
2	0	1
3	-1	-1

For the means model, the extension is straightforward:

Category Code

1	1	0	0
2	0	1	0
3	0	0	1

For multigroup ANOVA, the models have the same form as for the two-group ANOVA above. The corresponding hypotheses for testing whether there are differences between means are:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad (\text{effects model})$$

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad (\text{means model})$$

You do not need to know how to produce coding variables to do ANOVA. SYSTAT does this for you automatically. All you need is a single variable that contains different values for each group. SYSTAT translates these values into different codes. It is important to remember, however, that regression and analysis of variance are not fundamentally different models. They are both instances of the general linear model.

Factorial ANOVA

It is possible to have more than one categorical variable in ANOVA. When this happens, you code each categorical variable exactly the same way as you do with multi-group ANOVA. The coded design variables are then added as a full set of predictors in the model.



ANOVA factors can interact. For example, a treatment may enhance bar pressing by male rats, yet suppress bar pressing by female rats. To test for this possibility, you can add (to your model) variables that are the product of the *main effect* variables already coded. This is similar to what you do when you construct polynomial models. For example, this is a model without an interaction:

$$y = \text{CONSTANT} + \text{treat} + \text{sex}$$

This is a model that contains interaction:

$$y = \text{CONSTANT} + \text{treat} + \text{sex} + \text{treat} * \text{sex}$$

If the hypothesis test of the coefficients for the *TREAT*SEX* term is significant, then you must qualify your conclusions by referring to the interaction. You might say, "It works one way for males and another for females."

Data Screening and Assumptions

Most analyses have assumptions. If your data do not meet the necessary assumptions, then the resulting probabilities for the statistics may be suspect. Before an ANOVA, look for:

- Violations of the equal variance assumption. Your groups should have the same dispersion or spread (their shapes do not differ markedly).
- Symmetry. The mean of each group should fall roughly in the middle of the spread (the within-group distributions are not extremely skewed).
- Independence of the group means and standard deviations (the size of the group means is not related to the size of their standard deviations).
- Gross outliers (no values stand apart from the others in the batch).

Graphical displays are useful for checking assumptions. For analysis of variance, try dit plots, box-and-whisker displays, or bar charts with standard error bars.

Levene Test

Analysis of variance assumes that the data within cells are independent and normally distributed with equal variances. This is the ANOVA equivalent of the regression assumptions for residuals. When the homogeneous variance part of the assumptions is

false, it is sometimes possible to adjust the degrees of freedom to produce an approximately distributed *F-ratio*.

Levene (1960) proposed a test for unequal variances. You can use this test to determine whether you need an unequal variance *F* test. Simply fit your model in ANOVA and save residuals. Then transform the residuals into their absolute values. Merge these with your original grouping variable(s). Then redo your ANOVA on the absolute residuals. If it is significant, then you should consider using the separate variances test.

Before doing all this work, you should do a box plot by groups to see whether the distributions differ. If you see few differences in the spread of the boxes, Levene's test is unlikely to be significant.

Pairwise Mean Comparisons

The results in an ANOVA table serve only to indicate whether the means differ significantly or not. They do not indicate which mean differs from another.

To report the pairs of means that differ significantly, you might think of computing a two-sample *t* test for each pair; however, do *not* do this. The probability associated with the two-sample *t* test assumes that *only one test is performed*. When several means are tested pairwise, the probability of finding one significant difference by chance alone increases rapidly with the number of pairs. If you use a 0.05 significance level to test that means *A* and *B* are equal and to test that means *C* and *D* are equal, the overall acceptance region is now 0.95×0.95 , or 0.9025. Thus, the acceptance region for two independent comparisons carried out simultaneously is about 90%, and the critical region is 10% (instead of the desired 5%). For six pairs of means tested at the 0.05 significance level, the probability of a difference falling in the critical region is not 0.05 but $1 - (0.95)^6 = 0.265$. For 10 pairs, this probability increases to 0.40. The result of following such a strategy is to declare differences as significant when they are not.

As an alternative to the situation described above, SYSTAT provides fifteen techniques to perform pairwise mean comparisons. You have to choose a proper test

based on the variance assumptions and the error rate to be controlled. SYSTAT offers the following tests divided into two parts based on variance assumptions:

Equal Variance

Tukey
Bonferroni
Fisher's LSD
Sidak

Scheffé

Tukey's b
Duncan
Ryan-Einot-Gabriel-Welsch Q
Hochberg's GT2
Gabriel
Student-Newman-Keuls
Dunnett

Unequal Variance

Tamhane's T2
Games-Howell
Dunnett's T3

The Student-Newman-Keuls procedure (S-N-K) and Duncan's multiple range test control neither the individual nor the family-wise error rates. Duncan's test has been heavily criticized in the statistical literature; it gives many more statistically significant differences than is warranted and does not really protect the significance level. As a general rule, Fisher's LSD is one of the more liberal procedures (more likely to declare means different), but it does not control the family-wise error rate. Tukey's and Scheffé's methods are conservative, with Scheffé's method being more conservative than Tukey's method.

There is an abundance of literature covering multiple comparisons (see Miller, 1985); however, a few points are worth noting here:

- If you have a small number of groups, the Bonferroni pairwise procedure will often be more powerful (sensitive). For more groups, consider the Tukey method. Try all the methods in ANOVA (except Fisher's LSD) and pick the best one.
- Carrying out all possible pairwise comparisons is a waste of power. Think about a meaningful subset of comparisons and test this subset with Bonferroni levels. To do this, divide your critical level, say 0.05, by the number of comparisons you are making. You will almost always have more power than with any other pairwise multiple comparison procedure.
- Some popular multiple comparison procedures do not maintain their claimed protection levels. Other stepwise multiple range tests, such as the Student-

Newman-Keuls and Duncan's tests, have not been conclusively demonstrated to maintain overall protection levels for all possible distributions of means.

- Some tests produce and test homogeneous subsets of group means instead of testing each pair of the group means.
- Some tests come under unequal variance assumptions and use group variances instead of MSE to compare the group means.

Linear and Quadratic Contrasts

Contrasts are used to test relationships among means. A contrast is a linear combination of means μ_i with coefficients α_i :

$$\alpha_1\mu_1 + \alpha_2\mu_2 + \dots + \alpha_k\mu_k = 0$$

where $\alpha_1 + \alpha_2 + \dots + \alpha_k = 0$. In SYSTAT, hypotheses can be specified about contrasts and tests performed. Typically, the hypothesis has the form:

$$H_0: \alpha_1\mu_1 + \alpha_2\mu_2 + \dots + \alpha_k\mu_k = 0$$

The test statistic for a contrast is similar to that for a two-sample t test; the result of the contrast (a relation among means, such as mean A minus mean B) is in the numerator of the test statistic, and an estimate of within-group variability (the pooled variance estimate or the error term from the ANOVA) is part of the denominator.

You can select contrast coefficients to test:

- Pairwise comparisons (test for a difference between two particular means)
- A linear combination of means that are meaningful to the study at hand (compare two treatments versus a control mean)
- Linear, quadratic, or the like increases (decreases) across a set of ordered means (that is, you might test a linear increase in sales by comparing people with *no* training, those with *moderate* training, and those with *extensive* training)

Many experimental design texts place coefficients for linear and quadratic contrasts for three groups, four groups, and so on, in a table. SYSTAT allows you to type your contrasts or select a polynomial option. A polynomial contrast of order 1 is linear; of order 2, quadratic; of order 3, cubic; and so on.

Unbalanced Designs

An unbalanced factorial design occurs when the numbers of cases in cells are unequal and not proportional across rows or columns. The following is an example of a 2×2 design:

	<i>B1</i>	<i>B2</i>
<i>A1</i>	1	5
	2	3
		4
<i>A2</i>	6	2
	7	1
	9	5
	8	3
	4	

Unbalanced designs require a least-squares procedure like the General Linear Model because the usual maximum likelihood method of adding up the sum of squared deviations from cell means and the grand mean does not yield maximum likelihood estimates of effects. The General Linear Model adjusts for unbalanced designs when you get an ANOVA table to test hypotheses.

However, the estimates of effects in the unbalanced design are no longer orthogonal (and thus statistically independent) across factors and their interactions. This means that the sum of squares associated with one factor depends on the sum of squares for another or its interaction.

Analysts accustomed to using multiple regression have no problem with this situation because they assume that their independent variables in a model are correlated. Experimentalists, however, often have difficulty speaking of a main effect *conditioned* on another. Consequently, there is extensive literature on hypothesis testing methodology for unbalanced designs (for example, Speed and Hocking, 1976, and Speed, Hocking, and Hackney, 1978), and there is no consensus on how to test hypotheses with non-orthogonal designs.

Some statisticians advise you to do a series of hierarchical tests beginning with interactions. If the highest-order interactions are insignificant, drop them from the model and recompute the analysis. Then, examine the lower-order interactions. If they are insignificant, recompute the model with main effects only. Some computer programs automate this process and print sum of squares and *F* tests according to the hierarchy (ordering of effects) you specify in the model. These are often called Type I sum of squares.

This procedure is analogous to stepwise regression in which hierarchical subsets of models are tested. This example assumes that you have specified the following model:

$$Y = \text{CONSTANT} + a + b + c + a*b + a*c + b*c + a*b*c$$

The hierarchical approach tests the following models:

$$Y = \text{CONSTANT} + a + b + c + a*b + a*c + b*c + a*b*c$$

$$Y = \text{CONSTANT} + a + b + c + a*b + a*c + b*c$$

$$Y = \text{CONSTANT} + a + b + c + a*b + a*c$$

$$Y = \text{CONSTANT} + a + b + c + a*b$$

$$Y = \text{CONSTANT} + a + b + c$$

$$Y = \text{CONSTANT} + a + b$$

$$Y = \text{CONSTANT} + a$$

The problem with this approach, however, is that plausible subsets of effects are ignored if you examine only one hierarchy. The following model, which may be the best fit to the data, is never considered:

$$Y = \text{CONSTANT} + a + b + a*b$$

Furthermore, if you decide to examine all the other plausible subsets, you are really doing all possible subsets regression, and you should use Bonferroni confidence levels before rejecting a null hypothesis. The example above has 127 possible subset models (excluding ones without a CONSTANT). Interactive stepwise regression allows you to explore subset models under your control.

If you have done an experiment and have decided that higher-order effects (interactions) are of enough theoretical importance to include in your model, you should condition every test on all other effects in the model you selected. This is the classical approach of Fisher and Yates. It amounts to using the default F values on the ANOVA output, which are the same as the Type III sum of squares.

Probably the most important reason to stay with one model is that if you eliminate a series of effects that are not quite significant (for example, $p = 0.06$), you could end up with an incorrect subset model because of the dependencies among the sum of squares. In summary, if you want other sum of squares, compute them. You can supply the mean square error to customize sum of squares by using a hypothesis test in GLM, selecting MSE, and specifying the mean square error and degrees of freedom.

Repeated Measures

In factorial ANOVA designs, each subject is measured once. For example, the assumption of independence would be violated if a subject is measured first as a control group member and later as a treatment group member. However, in a repeated measures design, the same variable is measured several times for each subject (case). A paired-comparison t test is the most simple form of a repeated measures design (for example, each subject has a *before* and *after* measure).

Usually, it is not necessary for you to understand how SYSTAT carries out calculations; however, repeated measures is an exception. It is helpful to understand the quantities SYSTAT derives from your data. First, remember how to calculate a paired-comparison t test by hand:

- For each subject, compute the difference between the two measures.
- Calculate the average of the differences.
- Calculate the standard deviation of the differences.
- Calculate the test statistic using this mean and standard deviation.

SYSTAT derives similar values from your repeated measures and uses them in analysis-of-variance computations to test changes across the repeated measures (within subjects) as well as differences between groups of subjects (between subjects.) Tests of the within-subjects values are called polynomial tests of order 1, 2, ..., up to k , where k is one less than the number of repeated measures. The first polynomial is used to test linear changes (for example, do the repeated responses increase (or decrease) around a line with a significant slope?). The second polynomial tests if the responses fall along a *quadratic* curve, and so on.

For each case, SYSTAT uses **orthogonal contrast coefficients** to derive one number for each polynomial. For the coefficients of the linear polynomial, SYSTAT uses $(-1, 0, 1)$ when there are three measures; $(-3, -1, 1, 3)$ when there are four measures; and so on. When there are three repeated measures, SYSTAT multiplies the first by -1 , the second by 0 , and the third by 1 , and sums these products (this sum is then multiplied by a constant to make the sum of squares of the coefficients equal to 1). Notice that when the responses are the same, the result of the polynomial contrast is 0; when the responses fall closely along a line with a steep slope, the polynomial differs markedly from 0.

For the coefficients of the quadratic polynomial, SYSTAT uses $(1, -2, 1)$ when there are three measures; $(1, -1, -1, 1)$ when there are four measures; and so on. The cubic and higher-order polynomials are computed in a similar way.

Let us continue the discussion for a design with three repeated measures. Assume that you record body weight once a month for three months for rats grouped by diet. (Diet A includes a heavy concentration of alcohol and Diet B consists of normal lab chow.) For each rat, SYSTAT computes a linear component and a quadratic component. SYSTAT also sums weights to derive a *total* response. These derived values are used to compute two analysis of variance tables:

- The *total* response is used to test between-group differences; that is, the total is used as the dependent variable in the usual factorial ANOVA computations. In the example, this test compares total weight for Diet A against that for Diet B. This is analogous to a two-sample *t* test using total weight as the dependent variable.
- The linear and quadratic components are used to test changes across the repeated measures (within subjects) and also to test the interaction of the within factor with the grouping factor. If the test for the linear component is significant, you can report a significant linear increase in weight over the three months. If the test for the quadratic component is also significant (but much less so than the linear component), you might report that growth is predominantly linear, but there is a significant curve in the upward trend.
- A significant interaction between Diet (the between-group factor) and the linear component across time might indicate that the slopes for Diet A and Diet B differ. This test may be the most important one for the experiment.

Assumptions in Repeated Measures

SYSTAT computes both univariate and multivariate statistics. Like all standard ANOVA procedures, the univariate repeated measures approach requires that the distributions within cells be normal. The univariate repeated measures approach also requires that the covariances between all possible pairs of repeated measures be equal. (Actually, the requirement is slightly less restrictive, but this difference is of little practical importance.) Of course, the usual ANOVA requirement that all variances within cells are equal still applies; thus, the covariance matrix of the measures should have a constant diagonal and equal elements off the diagonal. This assumption is called **compound symmetry**.

The multivariate analysis does not require compound symmetry. It requires that the covariance matrices within groups (there is only one group in this example) be equivalent and that they be based on multivariate normal distributions. If the classical assumptions hold, then you should generally ignore the multivariate tests at the bottom

of the output and stay with the classical univariate ANOVA table because the multivariate tests will be generally less powerful.

There is a middle approach. The Greenhouse-Geisser and Huynh-Feldt statistics are used to adjust the probability for the classical univariate tests when compound symmetry fails. (Huynh-Feldt is a more recent adjustment to the conservative Greenhouse-Geisser statistic.) If the Huynh-Feldt *p-values* are substantially different from those under the column directly to the right of the *F-ratio*, then you should be aware that compound symmetry has failed. In this case, compare the adjusted *p-values* under Huynh-Feldt to those for the multivariate tests.

If all else fails, single degree-of-freedom polynomial tests can always be trusted. If there are several to examine, however, remember that you may want to use Bonferroni adjustments to the probabilities; that is, divide the normal value (for example, 0.05) by the number of polynomial tests you want to examine. You need to make a Bonferroni adjustment only if you are unable to use the summary univariate or multivariate tests to protect the overall level; otherwise, you can examine the polynomials without penalty if the overall test is significant.

See Timm (2002) for a discussion on repeated measures.

Issues in Repeated Measures Analysis

Repeated measures designs can be generated in SYSTAT with a single procedure. You need not worry about weighting cases in unbalanced designs or selecting error terms. The program does this automatically; however, you should keep the following in mind:

- The sum of squares for the univariate *F* tests are pooled across subjects within groups and their interactions with trials. This means that the traditional analysis method has highly restrictive assumptions. You must assume that the variances within cells are homogeneous and that the covariances across all pairs of cells are equivalent (compound symmetry). There are some mathematical exceptions to this requirement, but they rarely occur in practice. Furthermore, the compound symmetry assumption rarely holds for real data.
- Compound symmetry is *not* required for the validity of the single degree-of-freedom polynomial contrasts. These polynomials partition sum of squares into orthogonal components. You should routinely examine the magnitude of these sum of squares relative to the hypothesis sum of squares for the corresponding univariate repeated measures *F* test when your trials are ordered on a scale.
- Think of the repeated measures output as an expanded traditional ANOVA table.
 - The effects are printed in the same order as they appear in Winer, Brown and

Michels (1991) and other texts, but they include the single degree-of-freedom and multivariate tests to protect you from false conclusions. If you are satisfied that both are in agreement, you can delete the additional lines in the output file.

- You can test any hypothesis after you have estimated a repeated measures design and examined the output. For example, you can use polynomial contrasts to test single degree-of-freedom components in an unevenly spaced design. You can also use difference contrasts to do post hoc tests on adjacent trials.

SYSTAT's Sum of Squares

SYSTAT provides several types of sum of squares for testing hypotheses. The following names for these sum of squares are not statistical terms, but they were popularized originally by SAS GLM.

Type I. Type I sum of squares are adjusted for those terms which appear in the model after the term in question; some books refer to the method of obtaining this type of sum of squares as the hierarchical decomposition or the sequential sum of squares method (Milliken and Johnson, 2004). Type I sum of squares are computed from the difference between the residual sum of squares of two different models. The particular models needed for the computation depend on the order of the variables in the MODEL statement.

For example, if the model is:

```
MODEL y = CONSTANT + a + b + a*b
```

then the sum of squares for $A*B$ is produced from the difference between SSE (sum of squared error) in the two following models:

```
MODEL y = CONSTANT + a + b
MODEL y = CONSTANT + a + b + a*b
```

Similarly, the Type I sum of squares for B in this model is computed from the difference in SSE between the following models:

```
MODEL y = CONSTANT + a
MODEL y = CONSTANT + a + b
```

Finally, the Type I sum of squares for A is computed from the difference in residual sum of squares for the following:

```
MODEL y = CONSTANT
MODEL y = CONSTANT + a
```

In summary, to compute sum of squares, move from right to left and construct models which differ by the right-most term only. Type I sum of squares are commonly used for:

- A balanced ANOVA model where effects are specified in a hierarchical manner, viz., the main effect, first order interaction, second order interactions, and so on.
- A polynomial regression model in which terms in the model are ordered as per their degree.
- A purely nested model in which the effect is specified in the proper order.

Type II. Type II sum of squares is computed similarly to Type I except that main effects and interactions determine the ordering of differences instead of the MODEL statement order. For the above model, Type II sum of squares for the interaction is computed from the difference in residual sum of squares for the following models:

```
MODEL y = CONSTANT + a + b
MODEL y = CONSTANT + a + b + a*b
```

For the B effect, difference the following models:

```
MODEL y = CONSTANT + a + b
MODEL y = CONSTANT + a
```

For the A effect, difference the following (this is not the same as for Type I):

```
MODEL y = CONSTANT + a + b
MODEL y = CONSTANT + b
```

In summary, include interactions of the same order as well as all lower order interactions and main effects when differencing to get an interaction. When getting sum of squares for a main effect, difference against all other main effects and interactions involved with these main effects. The Type II sum of squares method is commonly used for:

- ANOVA model with unbalanced cell sizes (unbalanced ANOVA)
- ANOVA model that has main effects only
- Any regression model
- Nested design

Type III. Type III sum of squares are the default for ANOVA and are much simpler to understand. Simply difference from the full model, leaving out only the term in question. For example, the Type III sum of squares for A is taken from the following two models:

$$\begin{aligned}\text{MODEL } y &= \text{CONSTANT} + b + a*b \\ \text{MODEL } y &= \text{CONSTANT} + a + b + a*b\end{aligned}$$

The Type III sum of squares method is commonly used for:

- Any models in Type I and Type II
- Any balanced or unbalanced ANOVA
- Any ANOVA models with no missing cells

Type IV. Type IV sum of squares are designed for the missing cells designs and are not easily presented in the above terminology. They are produced by balancing over the means of nonmissing cells not included in the current hypothesis. SYSTAT has options to choose from three types of sum of squares, i.e. Type I, Type II, and Type III; you can choose one of these sum of squares for the analysis.

By default SYSTAT produces Type III sum of squares. The user should take care in choosing the appropriate choice for the sum of squares. There is often a strong temptation to choose the most significant sum of squares without understanding the hypothesis being tested.

Finally, Type IV is produced by the careful use of SPECIFY in testing means models. The advantage of this approach is that the user is always aware that sums of squares depend on explicit mathematical models rather than additions and subtractions of dimensionless quantities.

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. B. N. Petrov, and F. Csaki, eds. Second International Symposium on Information Theory. Budapest: Akademiai Kiado, pp. 267-281.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC 19, 716-723.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.
- Burnham, K.P., and Anderson, D.R. (2003). *Model selection and multimodel inference: A*

- practical information-theoretic approach*. 2nd ed. New York: Springer-Verlag.
- Flack, V. F. and Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, 41, 84–86.
- Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37, 152–155.
- Hocking, R. R. (1983). Developments in linear regression methodology: 1959–82. *Technometrics*, 25, 219–230.
- Hurvich, C.M. and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.
- Levene, H. (1960). Robust tests for equality of variance. I. Olkin, ed., *Contributions to Probability and Statistics*. Palo Alto, Calif.: Stanford University Press, 278–292.
- Lovell, M. C. (1983). Data Mining. *The Review of Economics and Statistics*, 65, 1–12.
- Miller, R. (1985). Multiple comparisons. Kotz, S. and Johnson, N. L., eds., *Encyclopedia of Statistical Sciences*, vol. 5. New York: John Wiley & Sons, 679–689.
- Milliken, G. A. and Johnson, D. E. (2004). *Analysis of messy data*, Vol. 1: *Designed Experiments*. 2nd ed. Boca Raton, FL: Chapman & Hall / CRC.
- Rencher, A. C. and Pun, F. C. (1980). Inflation of R-squared in best subset regression. *Technometrics*, 22, 49–54.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- Speed, F. M. and Hocking, R. R. (1976). The use of the $r()$ - notation with unbalanced data. *The American Statistician*, 30, 30–33.
- Speed, F. M., Hocking, R. R., and Hackney, O. P. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73, 105–112.
- Timm, N.H. (2002). *Applied multivariate analysis*. New York: Springer-Verlag.
- Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35, 234–242.
- Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, 86, 168–174.
- Wilkinson, L. and Dallal, G.E. (1982). Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics*, 24, 25–28.
- Winer, B. J., Brown, D. R., and Michels, K.M. (1991). *Statistical principles in experimental design*, 3rd ed. New York: McGraw-Hill.

Linear Models I: Linear Regression

Leland Wilkinson and Mark Coward

(revised by Soumyajit Ghosh and S.R.Kulkarni)

The model for simple linear regression is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where y is the dependent variable, x is the independent variable, and the β 's are the regression parameters (the intercept and the slope of the line of best fit). The model for multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$$

The Linear Regression feature offers three methods for fitting a multiple linear regression model: Least Squares Regression, Ridge Regression, and Bayesian Regression. Least Squares Regression estimates and tests simple and multiple linear regression models. The ability to do stepwise regression is available in three ways: use the default values, specify your own selection criteria, or at each step, interactively select a variable to add or remove from the model. SYSTAT offers three tests for checking normality: Kolmogorov-Smirnov, Lilliefors's test, Shapiro-Wilk test, and Anderson-Darling test, if opted. For each model you fit in Least Squares Regression, SYSTAT reports R^2 , adjusted R^2 , the standard error of the estimate, and an ANOVA table for assessing the fit of the model. AIC, AIC (Corrected) and Schwarz's BIC values are also provided for each fitted model. For more information on AIC and Schwarz's (1978) BIC refer to Chapter 1: Linear Models, "Variable Selection" on page 15 in *Statistics II*. For each variable in the model, the output includes the estimate of the regression coefficient, the standard error of the

coefficient, the standardized coefficient, tolerance, variance inflation factor (*VIF*), and a *t* statistic for measuring the usefulness of the variable in the model. A plot of residuals against the predicted values is provided. Also, in the case of single-predictor (independent variable), a plot of the fitted regression line with confidence limits for a single mean response and prediction limits for new observations is provided, and only fitted model in case of two predictors.

When the predictor variables are correlated, i.e. when multicollinearity exists, the least-squares estimates of regression coefficients tend to have a large sampling variability. In such a situation, ridge regression offers a method to obtain better estimates of regression coefficients. Two types of ridge coefficients: standardized coefficients and unstandardized coefficients are computed. A plot of the ridge factor against the ridge coefficients is also available. The technique of Partial Least Squares (PLS) regression can also be a remedy for multicollinearity. PLS reduces the dimensionality of the regression problem by using linear combinations of the predictors, in the process of which multicollinearity may be reduced or removed. For details of PLS, see "Partial Least Squares Regression" on page 357 in Statistics III.

Bayesian regression provides another paradigm for fitting a multiple linear regression model. The prior distribution for the regression parameters used in this feature is a (multivariate) Normal-Gamma distribution or a diffuse. Bayes estimates and credible intervals for the regression coefficients are computed. Also, the parameters of the posterior distribution are provided along with plots of prior and posterior densities of the regression coefficients.

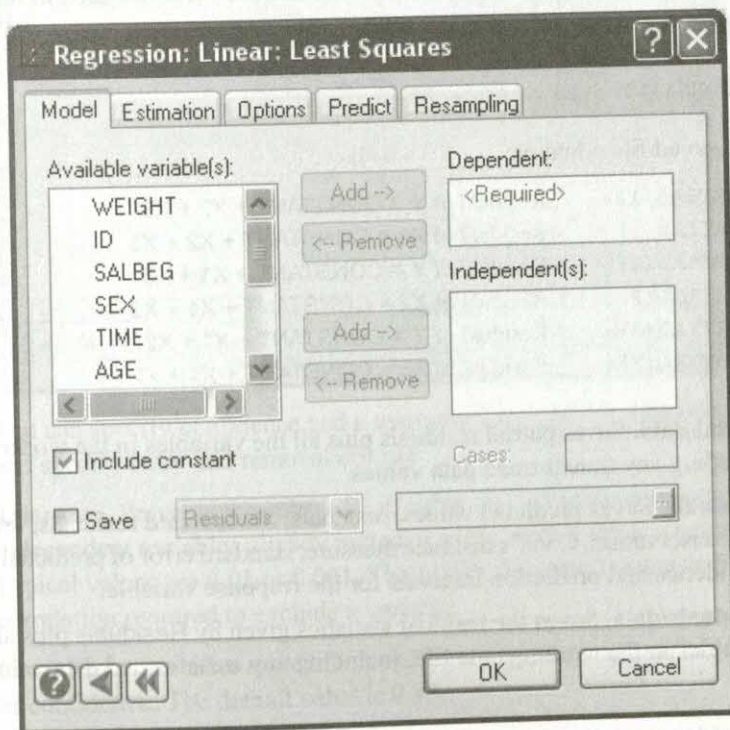
Resampling procedures are available only with Least Squares Regression. SYSTAT gives a summarization based on resampling for Linear Regression. You can get resampling-based estimates of the regression coefficients along with their bias and standard error. Under bootstrap, you will also get confidence intervals of coefficients using two popular methods, viz., Percentile method and Bias corrected and accelerated method.

Linear Regression in SYSTAT

Least Squares Regression Dialog Box

To open Least Squares Regression dialog box, from the menus choose:

Analyze
Regression
Linear
Least Squares...



The following options can be specified:

Include constant. Includes the constant in the regression equation. Deselect this option to remove the constant. You almost never want to remove the constant, and you should be familiar with no-constant regression terminology before considering it.

Cases. If your data are in the form of a correlation matrix, enter the number of cases used to compute the correlation matrix.

Save. You can save residuals and other data to a new data file. The following alternatives are available:

- **Adjusted.** Saves the adjusted estimates of the regression coefficients.
- **Adjusted/data.** Saves the adjusted estimates plus all the variables in the working data file.
- **Coefficients.** Saves the estimates of the regression coefficients.
- **Model.** Saves statistics given in Residuals and the variables used in the model.
- **Partial.** Saves partial residuals. Suppose your model is:

$$Y = \text{CONSTANT} + X1 + X2 + X3$$

The saved file contains:

YPARTIAL (1) : Residual of $Y = \text{CONSTANT} + X2 + X3$
 XPARTIAL (1) : Residual of $X1 = \text{CONSTANT} + X2 + X3$
 YPARTIAL (2) : Residual of $Y = \text{CONSTANT} + X1 + X3$
 XPARTIAL (2) : Residual of $X2 = \text{CONSTANT} + X1 + X3$
 YPARTIAL (3) : Residual of $Y = \text{CONSTANT} + X1 + X2$
 XPARTIAL (3) : Residual of $X3 = \text{CONSTANT} + X1 + X2$

- **Partial/data.** Saves partial residuals plus all the variables in the working data file, including any transformed data values.
- **Residuals.** Saves predicted values, residuals, Studentized residuals, leverage for each observation, Cook's distance measure, standard error of predicted values, and confidence and prediction intervals for the response variable.
- **Residuals/data.** Saves the residual statistics given by Residuals plus all the variables in the working data file, including any transformed data values.

Estimation

To specify the Estimation option, click the Estimation tab in the Least Squares Regression dialog box.

Stepwise options. The following alternatives are available for stepwise entry and removal:

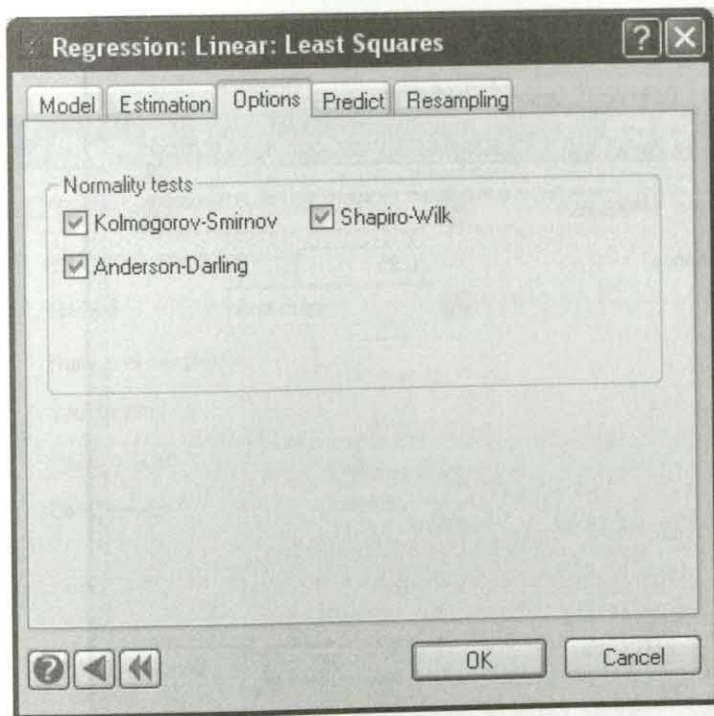
- **Backward.** Begins with all candidate variables in the model. At each step, SYSTAT removes the variable with the largest Remove value.
- **Forward.** Begins with no variables in the model. At each step, SYSTAT adds the variable with the smallest Enter value.
- **Automatic.** For Backward, at each step SYSTAT automatically removes a variable from your model. For Forward, SYSTAT automatically adds a variable to the model at each step.
- **Interactive.** At each step in the model building, you select the variable to enter or remove from the model.

You can also control the criteria used to enter and remove variables from the model:

- **Probability.** Specify probabilities to enter and to remove variable from the model. A variable is entered into the model if its alpha value is less than the specified Enter value and is removed from the model if its alpha value is greater than the specified Remove value. Specify values between 0 and 1.
- **F-ratio.** Specify F-to-enter and F-to-remove limits. Variables with *F-ratio* greater than the specified value are entered into the model if Tolerance permits and variables with *F-ratio* less than the specified value are removed from the model.
- **MaxStep.** Maximum number of steps.
- **Force.** Force the first *n* variables listed in your model to remain in the equation.

Options

To specify the options, click the Options tab in the Least Squares Regression dialog box.

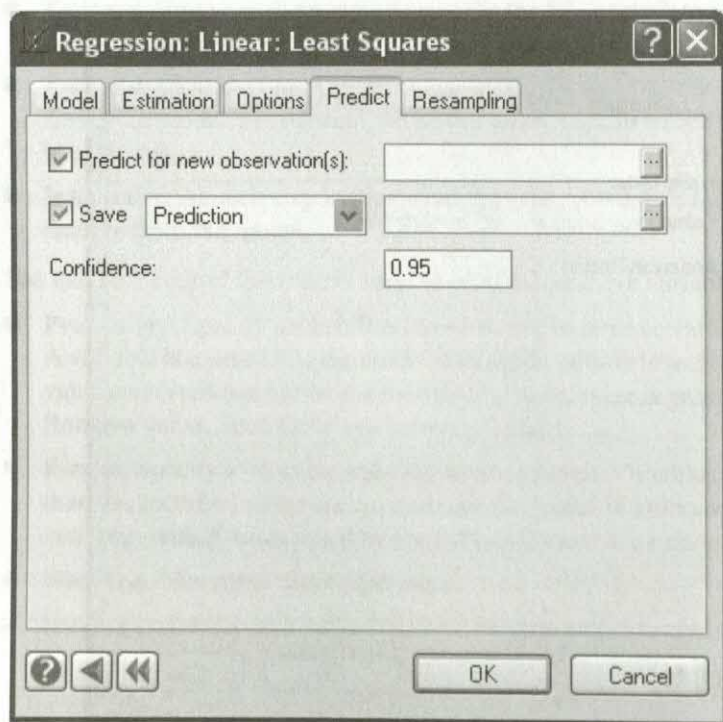


Normality tests. You can use the following tests to check the normality of residuals:

- **Kolmogorov-Smirnov.** It's a nonparametric test used for large samples. It is applied to continuous distributions and gives greater importance to the observations in the center than those at the tails.
- **Shapiro-Wilk.** The test provides Shapiro-Wilk test statistic and p-value for the residuals: the smaller the p-value, the worse is the fit.
- **Anderson-Darling.** Anderson-Darling test is a standard goodness of fit test. It gives greater importance to the observations in the tails than those at the center.

Predict

To predict the new values, click the Predict tab in the Least Squares regression dialog box.



The following options can be specified:

Prediction for new observation(s). Predicts the dependent variable value for given values of the predictors.

Confidence. Displays the confidence and prediction intervals at the desired level of confidence. The default value is 0.95.

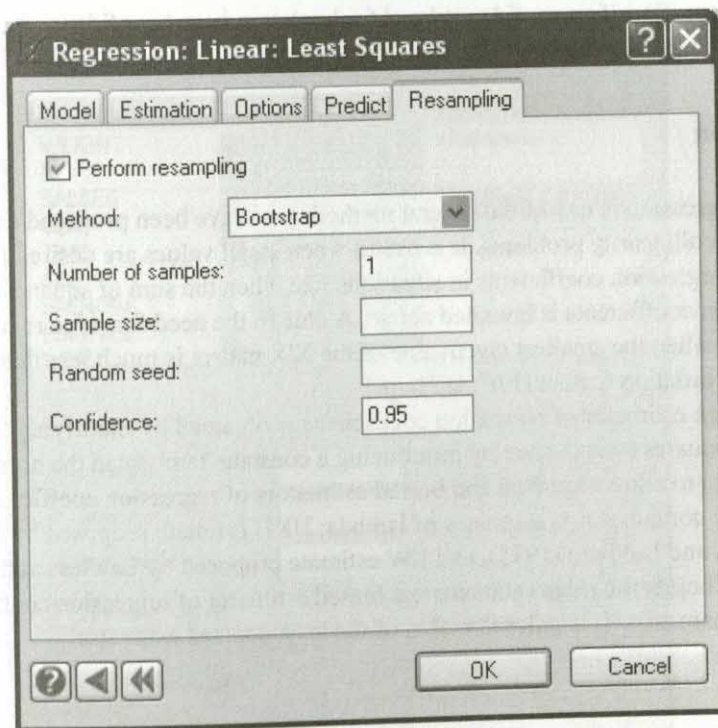
Save. You can save predicted values and new data on to a new data file. The following alternatives are available:

- **Prediction.** Saves the predicted values, standard errors of predicted values, lower and upper confidence limits of predicted values.

- **Prediction/New Data.** Saves the statistics given by Prediction plus variables in the model in new data file.

Resampling

Click the Resampling tab to specify different resampling options.



Perform resampling. Generates samples of cases and uses data thereof to carry out the same analysis on each sample.

Method. Three sampling methods are available:

- **Bootstrap.** Generates bootstrap samples. This is the default method.
- **Without replacement.** Generates subsamples without replacement.
- **Jackknife.** Generates jackknife samples.

Number of samples. Specify the number of samples to be generated. These samples are analyzed using the chosen method of sampling. The default is 1.

Sample size. Specify the size of each sample to be generated while resampling. The default sample size is the number of cases in the data file in use.

Random seed. Specify a random seed to be used while resampling. The default random seed is generated by the system.

Confidence. Specify a confidence level for bootstrap-based confidence interval. Enter any value between 0 and 1. The default is 0.95.

Ridge Regression

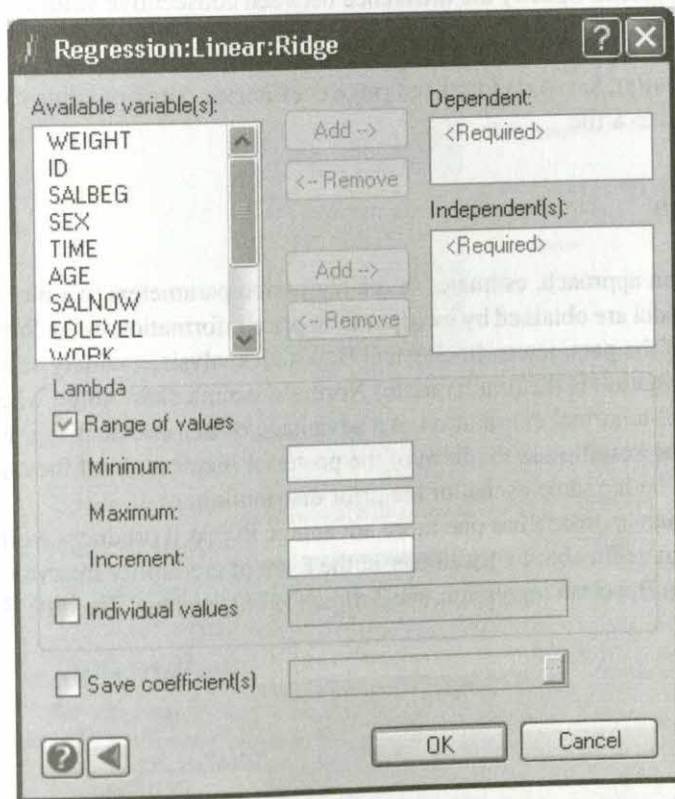
Ridge regression is one of the several methods that have been proposed as a remedy for multicollinearity problems. It is useful when small values are desired for the least-squares regression coefficients in situations like when the sum of squares of the regression coefficients is bounded above. A clue to the need for ridge regression is obtained when the smallest eigenvalue of the $X'X$ matrix is much less than 1 and the variance inflation factors (*VIF*) are large.

A ridge estimator of regression coefficients is obtained by modifying the method of least-squares (this is done by introducing a constant 'lambda' in the normal equations) to allow shrunken and biased estimators of regression coefficients. SYSTAT computes two estimates of lambda: HKB estimate proposed by Hoerl, Kennard, and Baldwin (1975), and LW estimate proposed by Lawless and Wang (1976). Though the ridge estimator is a biased estimator of regression coefficients, its mean square error is smaller than that of the least-squares estimator.

Ridge Regression Dialog Box

To open the Ridge Regression dialog box, from the main menus choose:

Analyze
Regression
Linear
Ridge...



Dependent. The variable to be predicted. The dependent variable should be quantitative in nature.

Independent(s). Select one or more variables. Normally, there exists high collinearity between the variables.

Lambda. You can specify the values of lambda to get HKB and LW estimates of optimal values of lambda.

You can specify individual lambda values or a range of lambda values to get the HKB and LW estimates of optimal values of lambda. Lambda is a real variable.

- **Range of values.** Specify a range of lambda values. The following options are provided for specifying the range of lambda values:
 - **Minimum.** Enter the minimum value or the start value of lambda.
 - **Maximum.** Enter the maximum value or the end value of lambda.
 - **Increment.** Specify the difference between consecutive values.
- **Individual values.** Specify desired set of lambda values.

Save coefficient(s). Saves standardized ridge coefficients corresponding to each of the lambda values to a file.

Bayesian Regression

In the Bayesian approach, estimates of the regression parameters in a multiple linear regression model are obtained by incorporating prior information in the form of a prior distribution of the parameters. In classical Bayesian analysis, a widely used choice of the prior distribution is the (multivariate) Normal-Gamma distribution when the error component has a normal distribution. An advantage of this choice is that it is a conjugate prior, resulting in the form of the posterior distribution of the regression parameters to be the same as that of the prior distribution.

The Bayesian approach has one more advantage in that it produces a direct probability statement about a parameter in the form of credibility intervals. For more information on Bayesian regression, see Zellner (1971), Box and Tiao (1973) and Press (1989).

Bayesian Regression Dialog Box

To obtain Bayesian Regression dialog box, from the menus choose:

Analyze
Regression
Linear
Bayesian...

Regression: Linear: Bayesian

Available variable(s):
WEIGHT
ID
SALBEG
SEX
TIME
AGE
SALNOW
EDLEVEL
WORK

Dependent: <Required>

Independent(s): <Required>

☐ Diffuse prior
☒ Normal-gamma prior

Normal prior parameters
Mean vector
☒ From keyboard
☐ From file

Covariance matrix
☒ From keyboard
☐ From file

Gamma prior parameters
Shape:
Scale:

☒ Include constant

Credibility: 0.95

☒ Save Coefficients

OK Cancel

Dependent. Select the variable you want to predict. The dependent variable should be continuous and numeric.

Independent. Select one or more independent variables.

Include constant. Includes the constant in the model (by default). Uncheck the box if you do not want to include the constant term in your regression equation.

Diffuse prior. Uses diffuse priors for estimation.

Normal-Gamma prior. Specify the Normal-Gamma conjugate priors for Bayesian estimation of regression coefficients.

- **Normal prior parameters.** Specify the parameters of the prior distribution of regression parameters.
- **Mean vector.** Enter the mean vector of the multivariate normal prior distribution of regression parameters either through the keyboard or using a file.
- **Covariance matrix.** Enter the covariance matrix of the multivariate normal prior distribution of regression parameters either through the keyboard or using a file.
- **Gamma prior parameters.** Enter the values of the scale and shape parameters of the gamma prior distribution for the inverse of the variance. The selection of gamma prior is optional. If one doesn't specify any gamma priors, only the regression coefficients of the posterior distribution are obtained.

Credibility. Enter the credibility coefficient (Bayesian analog of the confidence coefficient) to get the desired percentage credible interval. The default is 0.95.

Save. The following alternatives are available:

- **Coefficients.** Saves the estimates of the Bayesian regression coefficients to a specified file.
- **Residuals/data.** Saves all the predicted values, residuals and the original data.
- **Conditional covariance matrix.** Saves the conditional covariance matrix of Bayesian regression coefficients given sigma.
- **Marginal covariance matrix.** Saves the marginal covariance matrix of Bayesian regression coefficients.

Using Commands

For least squares regression

First, specify your data with `USE filename`. Continue with:

```
REGRESS
MODEL var=CONSTANT + var1 + var2 + ... / N=n
SAVE filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
WORK filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
ESTIMATE /MIX TOL=n NTEST = KS, SW, AD Quick NoQuick
SAVE filename / PREDICT NEWDATA
WORK filename / PREDICT NEWDATA
PREDICT filename /Confi=n Quick NoQuick
```

(use `START` instead of `ESTIMATE` for stepwise model building)

```
SAVE filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
START / FORWARD BACKWARD TOL=n ENTER=p REMOVE=p,
      FENTER=n FREMOVE=n FORCE=n
STEP / AUTO ENTER=p REMOVE=p FENTER=n FREMOVE=n
STOP/ Quick NoQuick
```

For getting the summarized resampling output, the following command should be given before the `ESTIMATE` command.

```
SAMPLE BOOT(m,n) or SIMPLE(m,n) or JACK / CONFI = c
```

For ridge regression

Select a data file using `USE filename` and continue with:

```
RIDGEREG
MODEL var = CONSTANT + var1 + var2 +...+ varn
SAVE filename
WORK filename
ESTIMATE / LMIN=a LMAX=b LSTEP=c or LAMBDA=11, 12,..., 1k
```


For Bayesian regression

```

BAYESIAN
MODEL var = CONSTANT + var1 + var2 +...+ varn
SAVE filename/COEFFICIENTS or RESIDUALS, DATA or CONDITIONAL
               or MARGINAL
WORK filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
ESTIMATE / MEAN = [b] or 'filename1' VAR = [v] or 'filename2'
SCALE=a SHAPE=c CREDIBILITY=d

```

Usage Considerations

Types of data. REGRESS uses the usual cases-by-variables data file or a covariance, correlation, or sum of squares and cross products matrix. Using matrix input requires specification of the sample size, which generated the matrix. RIDGEREG and BAYESIAN use rectangular data only.

Print options. For REGRESS, using PLENGTH MEDIUM, the output includes eigenvalues of $X'X$, condition indices, and variance proportions. PLENGTH LONG adds the correlation matrix of the regression coefficients to this output. For RIDGEREG and BAYESIAN regression, the output is standard for all PLENGTH options.

Quick Graphs. REGRESS plots the residuals against the predicted values. Also plots confidence limits for a single mean response and prediction limits for new observations in the single-predictor case for both original and new observations. And in case of two predictors it plots a fitted model. RIDGEREG plots a graph between the ridge factor and the ridge coefficients. BAYESIAN produces plots of the prior and the posterior densities of each regression coefficient and of the variance.

Saving files. REGRESS saves the results of the analysis (predicted values, residuals, confidence intervals, prediction intervals and diagnostics that identify unusual cases). RIDGEREG saves the ridge coefficients and BAYESIAN saves the estimates of the regression coefficients, residuals marginal and conditional covariance matrix.

BY groups. REGRESS, RIDGEREG, and BAYESIAN analyze data by groups.

Case frequencies. REGRESS, RIDGEREG, and BAYESIAN use the FREQ variable to duplicate cases. This inflates the degrees of freedom to be the sum of frequencies.

Case weights. REGRESS, RIDGEREG and BAYESIAN weight cases using the WEIGHT variable for rectangular data. You can perform cross-validation if the weight variable is binary and coded 0 or 1. SYSTAT computes predicted values for cases with zero weight even though they are not used to estimate the regression parameters.

Examples

Example 1 Simple Linear Regression

In this example, we explore the relation between gross domestic product per capita (GDP_CAP) and spending on the military (MIL) for 57 countries that report this information to the United Nations—we want to determine whether a measure of the financial well being of a country is useful for predicting its military expenditures. Our model is:

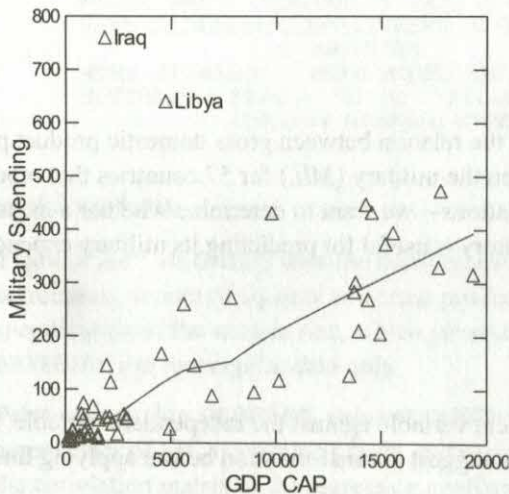
$$mil = \beta_0 + \beta_1 gdp_cap + \varepsilon$$

Initially, we plot the dependent variable against the independent variable. Such a plot may reveal outlying cases or suggest a transformation before applying linear regression.

The input is:

```
USE OURWORLD  
IF COUNTRY$='IRAQ' or COUNTRY$='LIBYA' THEN LET NAME$=COUNTRY$  
PLOT MIL*GDP_CAP / SMOOTH=LOWESS TENSION =0.500,  
YLABEL='Military Spending',  
SYMBOL=4 SIZE= 1.500 LABEL=NAME$,  
CSIZE=2.000
```

The output is:



To obtain the scatterplot, we created a new variable, *NAME\$*, that had missing values for all countries except Libya and Iraq. We then used the new variable to label plot points.

Iraq and Libya stand apart from the other countries—they spend considerably more for the military than countries with similar *GDP_CAP* values. The smoother indicates that the relationship between the two variables is fairly linear. Distressing, however, is the fact that many points clump in the lower left corner. Many data analysts would want to study the data after log-transforming both variables. We do this in another example, but now we estimate the coefficients for the data as recorded.

The input is:

```
REGRESS
  USE OURWORLD
  ID COUNTRY$
  MODEL MIL = CONSTANT + GDP_CAP
  ESTIMATE
```

The output is:

1 case(s) are deleted due to missing data.

Eigenvalues of Unit Scaled X'X

1	2
1.681	0.319

Condition Indices

1	2
1.000	2.294

Variance Proportions

	1	2
CONSTANT	0.160	0.840
GDP_CAP	0.160	0.840

Dependent Variable	MIL
N	56
Multiple R	0.646
Squared Multiple R	0.417
Adjusted Squared Multiple R	0.407
Standard Error of Estimate	136.154

Regression Coefficients B = (X'X)⁻¹X'Y

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	41.857	24.838	0.000	.	1.685
GDP_CAP	0.019	0.003	0.646	1.000	6.220

Regression Coefficients B = (X'X)⁻¹X'Y (contd...)

Effect	p-value
CONSTANT	0.098
GDP_CAP	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	41.857	-7.940	91.654	.
GDP_CAP	0.019	0.013	0.025	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	717100.891	1	717100.891	38.683	0.000
Residual	1.001E+006	54	18537.876		

*** WARNING *** :

Case Iraq is an Outlier (Studentized Residual : 6.956)

Case Libya is an Outlier (Studentized Residual : 4.348)

Durbin-Watson D Statistic : 2.046

First Order Autocorrelation : -0.032

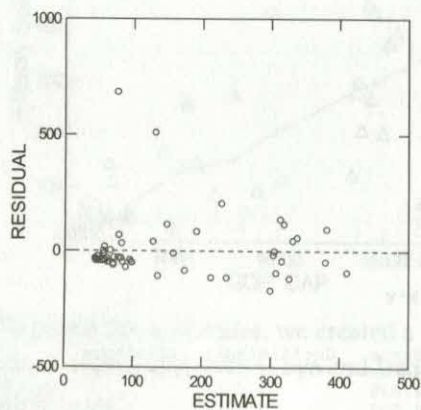
Information Criteria

AIC : 713.229

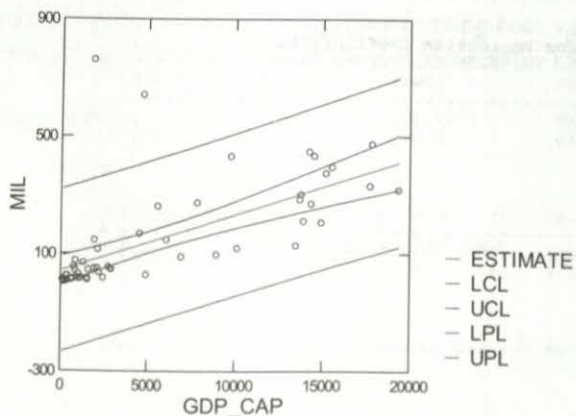
AIC (Corrected) : 713.690

Schwarz's BIC : 719.305

Plot of Residuals vs Predicted Values



Confidence Interval and Prediction Interval



SYSTAT reports that data are missing for one case. In the next line, it reports that 56 cases are used ($N = 56$). In the regression calculations, SYSTAT uses only the cases that have complete data for the variables in the model. However, when only the dependent variable is missing, SYSTAT computes a predicted value, its standard error, and a leverage diagnostic for the case. In this sample, Afghanistan did not report military spending.

When there is only one independent variable, *Multiple R* (0.646) is the simple correlation between *MIL* and *GDP_CAP*. *Squared multiple R* (0.417) is the square of this value, and it is the proportion of the total variation in the military expenditures accounted for by *GDP_CAP* (*GDP_CAP* explains 41.7% of the variability of *MIL*). Use Sum-of-Squares (*SS*) in the analysis of variance table to compute it:

$$717100.891 / (717100.891 + 1001045.288)$$

Adjusted squared multiple R is of interest for models with more than one independent variable. *Standard error of estimate* (136.154) is the square root of the residual mean square (18537.876) in the ANOVA table.

The estimates of the regression coefficients are 41.857 and 0.019, so the equation is:

$$\text{mil} = 41.857 + 0.019 * \text{gdp_cap}$$

The *Standard Error* of the estimated coefficients are in the next column and the standardized coefficients (*Std Coefficient*) follow. The latter are called beta weights by some social scientists. *Tolerance* and Variance inflation factor (*VIF*) are not relevant when there is only one predictor.

Next are *t* statistics (*t*)—the first (1.685) tests the significance of the difference of the constant from 0 and the second (6.220) tests the significance of the slope, which is equivalent to testing the significance of the correlation between military spending and *GDP_CAP*.

F-ratio in the analysis of variance table is used to test the hypothesis that the slope is 0 (or, for multiple regression, that all slopes are 0). The *F-ratio* is large when the independent variable(s) helps to explain the variation in the dependent variable. Here, there is a significant linear relation between military spending and *GDP_CAP*. Thus, we reject the hypothesis that the slope of the regression line is zero (*F-ratio* = 38.683, *p-value* < 0.0005).

It appears from the results above that *GDP_CAP* is useful for predicting spending on the military—that is, countries that are financially sound tend to spend more on the military than poorer nations. These numbers, however, do not provide the complete picture. Notice that SYSTAT warns us that the two countries (Iraq and Libya) with

unusual values could be distorting the results. We recommend that you consider transforming the data and that you save the residuals and other diagnostic statistics.

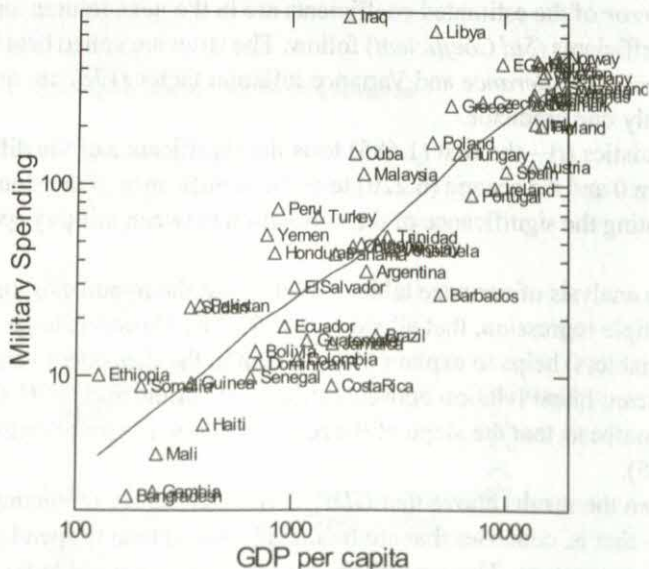
Example 2 Transformations

The data in the scatterplot in the simple linear regression example are not well suited for linear regression, as the heavy concentration of points in the lower left corner of the graph shows. Here are the same data plotted in log units.

The input is:

```
REGRESS
USE OURWORLD
PLOT MIL*GDP_CAP / SMOOTH=LOWESS TENSION =0.500,
XLABEL='GDP per capita',
XLOG=10 YLABEL='Military Spending' YLOG=10,
SYMBOL=4,2,3,SIZE= 1.250 LABEL=COUNTRY$,
CSIZE=1.450
```

The output is:



Except possibly for Iraq and Libya, the configuration of these points is better for linear modeling than that for the untransformed data.

We now transform both the y and x variables and refit the model, the input is:

```
REGRESS
  USE OURWORLD
  LET LOG_MIL = L10(MIL)
  LET LOG_GDP = L10(GDP_CAP)
  MODEL LOG_MIL = CONSTANT + LOG_GDP
  ESTIMATE
```

The output is:

1 case(s) are deleted due to missing data.

Eigenvalues of Unit Scaled $X'X$

1	2
1.984	0.016

Condition Indices

1	2
1.000	11.005

Variance Proportions

	1	2
CONSTANT	0.008	0.992
LOG_GDP	0.008	0.992

Dependent Variable	LOG_MIL
N	56
Multiple R	0.857
Squared Multiple R	0.734
Adjusted Squared Multiple R	0.729
Standard Error of Estimate	0.346

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	-1.308	0.257	0.000	.	-5.091
LOG_GDP	0.909	0.075	0.857	1.000	12.201

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.000
LOG_GDP	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	-1.308	-1.822	-0.793	.
LOG_GDP	0.909	0.760	1.058	1.000

Correlation Matrix of Regression Coefficients

	CONSTANT	LOG_GDP
CONSTANT	1.000	
LOG_GDP	-0.984	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	17.868	1	17.868	148.876	0.000
Residual	6.481	54	0.120		

*** WARNING *** :

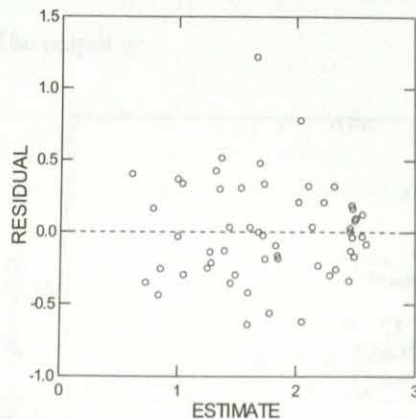
Case 22 is an Outlier (Studentized Residual : 4.004)

Durbin-Watson D Statistic : 1.810
First Order Autocorrelation : 0.070

Information Criteria

AIC : 44.160
AIC (Corrected) : 44.621
Schwarz's BIC : 50.236

Plot of Residuals vs Predicted Values



The *Squared multiple R* for the variables in log units is 0.734 (versus 0.417 for the untransformed values). That is, we have gone from explaining 41.7% of the variability of military spending to 73.4% by using the log transformations. The *F-ratio* is now 148.876—it was 38.683. Notice that we now have only one outlier (Iraq).

The Calculator

But what is the estimated model now?

$$\text{LOG_MIL} = -1.308 + 0.909 * \text{LOG_GDP}$$

However, many people don't think in "log units." Let's transform this equation (exponentiate each side of the equation):

$$10^{\log_mil} = 10^{(-1.308 + 0.909 * \log_gdp)}$$

$$mil = 10^{-1.308 + 0.909 * \log(gdp)}$$

$$mil = 10^{-1.308} * 10^{0.909 * \log(gdp)}$$

$$mil = 0.049 * (gdp_cap)^{0.909}$$

We used the calculator to compute 0.049. Type:

CALC $10^{(1.308)}$

and SYSTAT returns 0.049.

Example 3

Residuals and Diagnostics for Simple Linear Regression

In this example, we continue with the transformations example and save the residuals and diagnostics along with the data. Using the saved statistics, we create stem-and-leaf plots of the residuals and Studentized residuals. In addition, let's plot the Studentized residuals (to identify outliers in the y space) against leverage (to identify outliers in the x space) and use Cook's distance measure to scale the size of each plot symbol. In a second plot, we display the corresponding country names.

The input is:

```
REGRESS
USE OURWORLD
LET LOG_MIL = L10(MIL)
LET LOG_GDP = L10(GDP_CAP)
MODEL LOG_MIL = CONSTANT + LOG_GDP
SAVE MYRESULT / DATA RESID
ESTIMATE
USE MYRESULT
CLSTEM RESIDUAL STUDENT
PLOT STUDENT*LEVERAGE / SYMBOL=4,2,3 SIZE=cook
PLOT STUDENT*LEVERAGE / LABEL=COUNTRY$ SYMBOL=4,2,3
```

The output is:

Stem and Leaf Plot of Variable:
RESIDUAL, N = 56

Minimum : -0.644
Lower Hinge : -0.246
Median : -0.031
Upper Hinge : 0.203
Maximum : 1.216

```

-6 42
-5 6
-4 42
-3 554000
-2 H 65531
-1 9876433
-0 M 98433200
0 222379
1 1558
2 H 009
3 0113369
4 27
5 1
6
7 7

```

OutsideValues
12 1

1 Cases with missing values excluded from plot

Stem and Leaf Plot of Variable:
STUDENT, N = 56

Minimum : -1.923
Lower Hinge : -0.719
Median : -0.091
Upper Hinge : 0.591
Maximum : 4.004

```

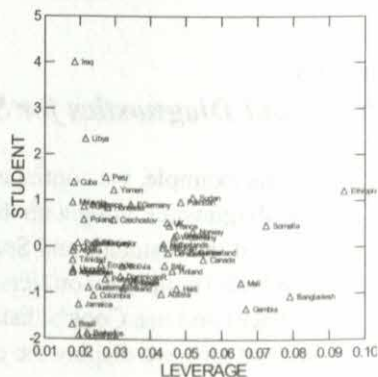
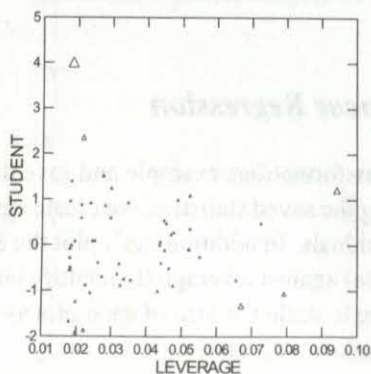
-1 986
-1 32000
-0 H 88877766555
-0 M 443322111000
0 M 000022344
0 H 555889999
1 0223
1 5
2 3

```

OutsideValues

4 0

1 Cases with missing values excluded from plot.



In the stem-and-leaf plots, Iraq's residual is 1.216 and is identified as an *Outside Value*. The value of its Studentized residual is 4.004, which is very extreme for the t distribution.

The case with the most influence on the estimates of the regression coefficients stands out at the top left (that is, it has the largest plot symbol). From the second plot, we identify this country as Iraq. Its value of Cook's distance measure is large because its Studentized residual is extreme. On the other hand, Ethiopia (furthest to the right), the case with the next most influence, has a large value of Cook's distance because its

value of leverage is large. Gambia has the third largest Cook value, and Libya, the fourth.

Deleting an Outlier and Normality Testing

Residual plots identify Iraq as the case with the greatest influence on the estimated coefficients. Let's remove this case from the analysis and check SYSTAT's warnings.

The input is:

```
REGRESS
  USE OURWORLD
  LET LOG_MIL = L10(MIL)
  LET LOG_GDP = L10(GDP_CAP)
  SELECT MIL < 700
  MODEL LOG_MIL = CONSTANT + LOG_GDP
  ESTIMATE/NTEST = KS, SW, AD
  SELECT
```

The output is:

Dependent Variable	LOG_MIL
N	55
Multiple R	0.886
Squared Multiple R	0.785
Adjusted Squared Multiple R	0.781
Standard Error of Estimate	0.306

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Coefficient	Tolerance	t
CONSTANT	-1.353	0.227	0.000	.	-5.949
LOG_GDP	0.916	0.066	0.886	1.000	13.896

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.000
LOG_GDP	0.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	18.129	1	18.129	193.107	0.000
Residual	4.976	53	0.094		

Test for Normality

	Test Statistic	p-value
K-S Test (Lilliefors)	0.071	0.669
Shapiro-Wilk Test	0.988	0.864
Anderson-Darling Test	0.224	>0.15


```
Durbin-Watson D Statistic | 1.763
First Order Autocorrelation | 0.086
```

Information Criteria

```
AIC | 29.931
AIC (Corrected) | 30.401
Schwarz's BIC | 35.953
```

Now there are no warnings about outliers. From the above results of normality tests, the assumption of residuals to be normal is satisfied.

Printing Residuals and Diagnostics

Let's look at some of the values in the *MYRESULT* file. We use the country name as the ID variable for the listing.

The input is:

```
USE MYRESULT
IDVAR COUNTRY$
FORMAT 10 3
LIST COOK LEVERAGE STUDENT MIL GDP_CAP
```

The output is:

Case ID	COOK	LEVERAGE	STUDENT	MIL	GDP_CAP
Ireland	0.013	0.032	-0.891	95.833	8970.885
Austria	0.023	0.043	-1.011	127.237	13500.299
Belgium	0.000	0.044	-0.001	283.939	13724.502
Denmark	0.000	0.045	-0.119	269.608	14363.064
(etc.)					
Libya	0.056	0.022	2.348	640.513	4738.055
Somalia	0.009	0.072	0.473	8.846	201.798
Afghanistan	.	0.075	.	.	189.128
(etc.)					

The value of *MIL* for Afghanistan is missing, so Cook's distance measure and Studentized residuals are not available (periods are inserted for these values in the listing).

Example 4

Multiple Linear Regression

In this example, we build a multiple regression model to predict total employment using values of six independent variables. The data were originally used by Longley (1967) to test the robustness of least-squares packages to multicollinearity and other sources of ill-conditioning. SYSTAT can print the estimates of the regression coefficients with more "correct" digits than the solution provided by Longley himself if you adjust the number of decimal places. By default, the first three digits after the decimal are displayed. After the output is displayed, you can use General Linear Model to test hypotheses involving linear combinations of regression coefficients.

The input is:

```
REGRESS
USE LONGLEY
PLENGTH LONG
MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
               ARMFORCE + POPULATN + TIME
ESTIMATE
```

The output is:

Eigenvalues of Unit Scaled X'X

1	2	3	4	5
6.861	0.082	0.046	0.011	0.000

Eigenvalues of Unit Scaled X'X

6	7
0.000	0.000

Condition Indices

1	2	3	4	5
1.000	9.142	12.256	25.337	230.424

Condition Indices

6	7
1048.080	43275.047

Variance Proportions

	1	2	3	4	5
CONSTANT	0.000	0.000	0.000	0.000	0.000
DEFLATOR	0.000	0.000	0.000	0.000	0.457
GNP	0.000	0.000	0.000	0.001	0.016
UNEMPLOY	0.000	0.014	0.001	0.065	0.006
ARMFORCE	0.000	0.092	0.064	0.427	0.115
POPULATN	0.000	0.000	0.000	0.000	0.010
TIME	0.000	0.000	0.000	0.000	0.000

Variance Proportions

	6	7
CONSTANT	0.000	1.000
DEFLATOR	0.505	0.038
GNP	0.328	0.655
UNEMPLOY	0.225	0.689
ARMFORCE	0.000	0.302
POPULATN	0.831	0.160
TIME	0.000	1.000

Dependent Variable	TOTAL
N	16
Multiple R	0.998
Squared Multiple R	0.995
Adjusted Squared Multiple R	0.992
Standard Error of Estimate	304.854

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	-3.482E+006	890420.384	0.000	.	-3.911
DEFLATOR	15.062	84.915	0.046	0.007	0.177
GNP	-0.036	0.033	-1.014	0.001	-1.070
UNEMPLOY	-2.020	0.488	-0.538	0.030	-4.136
ARMFORCE	-1.033	0.214	-0.205	0.279	-4.822
POPULATN	-0.051	0.226	-0.101	0.003	-0.226
TIME	1829.151	455.478	2.480	0.001	4.016

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.004
DEFLATOR	0.863
GNP	0.313
UNEMPLOY	0.003
ARMFORCE	0.001
POPULATN	0.826
TIME	0.003

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	-3.482E+006	-5.497E+006	-1.468E+006	.
DEFLATOR	15.062	-177.029	207.153	135.532
GNP	-0.036	-0.112	0.040	1788.513
UNEMPLOY	-2.020	-3.125	-0.915	33.619
ARMFORCE	-1.033	-1.518	-0.549	3.589
POPULATN	-0.051	-0.563	0.460	399.151
TIME	1829.151	798.788	2859.515	758.981

Correlation Matrix of Regression Coefficients

	CONSTANT	DEFLATOR	GNP	UNEMPLOY	ARMFORCE
CONSTANT	1.000				
DEFLATOR	-0.205	1.000			
GNP	0.816	-0.649	1.000		
UNEMPLOY	0.836	-0.555	0.946	1.000	
ARMFORCE	0.550	-0.349	0.469	0.619	1.000
POPULATN	-0.411	0.659	-0.833	-0.758	-0.189
TIME	-1.000	0.186	-0.802	-0.824	-0.549

Correlation Matrix of Regression Coefficients

	POPULATN	TIME
POPULATN	1.000	
TIME	0.388	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	1.842E+008	6	3.070E+007	330.285	0.000
Residual	836424.056	9	92936.006		

Durbin-Watson D Statistic | 2.559

First Order Autocorrelation | -0.348

Information Criteria

AIC	235.235
AIC (Corrected)	255.806
Schwarz's BIC	241.416

SYSTAT computes the eigenvalues by scaling the columns of the \mathbf{X} matrix so that the diagonal elements of $\mathbf{X}'\mathbf{X}$ are 1's and then factoring the $\mathbf{X}'\mathbf{X}$ matrix. In this example, most of the eigenvalues of $\mathbf{X}'\mathbf{X}$ are nearly 0, showing that the predictor variables comprise a relatively redundant set.

Condition indices are the square roots of the ratios of the largest eigenvalue to each successive eigenvalue. A condition index greater than 15 indicates a possible problem, and an index greater than 30 suggests a serious problem with collinearity (Belsley, Kuh, and Welsh, 1980). The condition indices in the Longley example show a tremendous collinearity problem.

Variance proportions are the proportions of the variance of the estimates accounted for by each principal component associated with each of the above eigenvalues. You should begin to worry about collinearity when a component associated with a high condition index contributes substantially to the variance of two or more variables. This is certainly the case with the last component of the Longley data. *TIME*, *GNP*, and *UNEMPLOY* load highly on this component. See Belsley, Kuh, and Welsh (1980) for more information about these diagnostics.

Adjusted squared multiple R is 0.992. The formula for this statistic is:

$$\text{adj. sq. multiple } R = R^2 - \frac{(p-1)}{(n-p)} * (1 - R^2)$$

where n is the number of cases and p is the number of predictors, including the constant.

Notice the extremely small tolerances in the output. *Tolerance* is 1 minus the multiple correlation between a predictor and the remaining predictors in the model.

The variance inflation factor (*VIF*) measures how much the variances of the estimated regression coefficient are inflated i.e., it identifies the independent variable with substantial multicollinearity with other independent variables. It is also defined as the reciprocal of *tolerance*.

$$VIF_j = \frac{1}{TOL} = \frac{1}{1 - R_j^2}$$

R_j^2 is the multiple correlation between a predictor and the remaining predictors in the model. If one of the regressor variables is nearly linearly dependent on some other regressors then R_j^2 will be near unity which implies tolerance to be close to zero and *VIFs* to be large. Large *VIFs* imply serious problems with multicollinearity. These tolerances and *VIFs* signal that the predictor variables are highly intercorrelated—a worrisome situation. This multicollinearity can inflate the standard errors of the coefficients, thereby attenuating the associated *F-ratio*, and can threaten computational accuracy.

Finally, SYSTAT produces the *Correlation matrix of regression coefficients*. In the Longley data, these estimates are highly correlated, further indicating that there are too many correlated predictors in the equation to provide stable estimates.

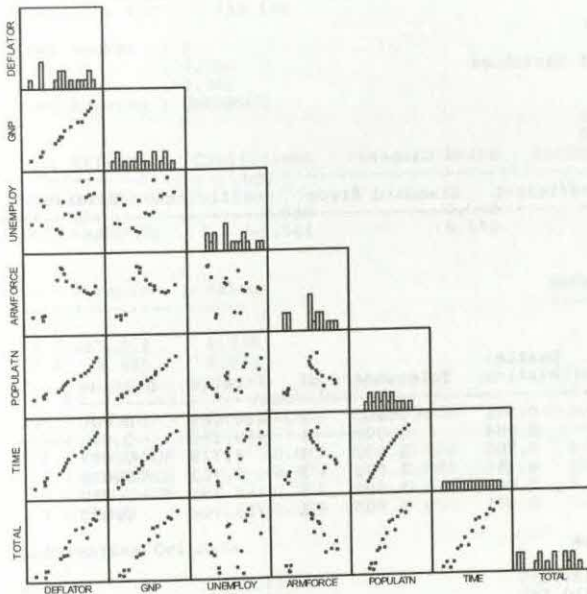
Scatterplot Matrix

Examining a scatterplot matrix of the variables in the model is often a beneficial first step in any multiple regression analysis. Nonlinear relationships and correlated predictors, both of which cause problems for multiple linear regression, can be uncovered before fitting the model.

The input is:

```
USE LONGLEY
SPLOM DEFLATOR GNP UNEMPLOY ARMFORCE POPULATN TIME TOTAL / HALF,
DENSITY=HIST
```

The output is:



Notice the severely nonlinear relationships of *ARMFORCE* with the other variables, as well as the near perfect correlations among several of the predictors. There is also a sharp discontinuity between post-war and 1950's behavior on *ARMFORCE*.

Example 5 Automatic Stepwise Regression

The following is an example of forward automatic stepping using the *LONGLEY* data.

The input is:

```

REGRESS
  USE LONGLEY
  MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
                ARMFORCE + POPULATN + TIME
  START / FORWARD
  STEP / AUTO
  STOP

```

The output is:

Stepwise Selection of Variables

```

Step Number : 0
R            : 0.000
R-square     : 0.000

```

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					

In	F-ratio	p-value
1		

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.971	1.000	1	230.089	0.000
3	GNP	0.984	1.000	1	415.103	0.000
4	UNEMPLOY	0.502	1.000	1	4.729	0.047
5	ARMFORCE	0.457	1.000	1	3.702	0.075
6	POPULATN	0.960	1.000	1	166.296	0.000
7	TIME	0.971	1.000	1	233.704	0.000

Information Criteria

```

AIC           : 309.619
AIC (Corrected) : 310.542
Schwarz's BIC : 311.164

```

```

Dependent Variable      : TOTAL
Minimum Tolerance for Entry into Model : 0.000
Forward Stepwise with Alpha-to-Enter : 0.150
Forward Stepwise with Alpha-to-Remove : 0.150

```

```

Step Number : 1
R           : 0.984
R-square    : 0.967
Term Entered : GNP

```

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	0.035	0.002	0.984	1.000	1

In	F-ratio	p-value
1		
3	415.103	0.000

Linear Models I: Linear Regression

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	-0.187	0.017	1	0.473	0.504
4	UNEMPLOY	-0.638	0.635	1	8.925	0.010
5	ARMFORCE	0.113	0.801	1	0.167	0.689
6	POPULATN	-0.598	0.018	1	7.254	0.018
7	TIME	-0.432	0.009	1	2.979	0.108

Information Criteria

AIC : 256.857
 AIC (Corrected) : 258.857
 Schwarz's BIC : 259.175

Step Number : 2
 R : 0.990
 R-square : 0.981
 Term Entered : UNEMPLOY

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	0.038	0.002	1.071	0.635	1
4	UNEMPLOY	-0.544	0.182	-0.145	0.635	1

In	F-ratio	p-value
1		
3	489.314	0.000
4	8.925	0.010

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	-0.073	0.016	1	0.064	0.805
5	ARMFORCE	-0.479	0.486	1	3.580	0.083
6	POPULATN	-0.164	0.006	1	0.334	0.574
7	TIME	0.308	0.002	1	1.259	0.284

Information Criteria

AIC : 250.494
 AIC (Corrected) : 254.131
 Schwarz's BIC : 253.585

Step Number : 3
 R : 0.993
 R-square : 0.985
 Term Entered : ARMFORCE

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	0.041	0.002	1.154	0.318	1
4	UNEMPLOY	-0.797	0.213	-0.212	0.385	1
5	ARMFORCE	-0.483	0.255	-0.096	0.486	1

In	F-ratio	p-value
1		
3	341.684	0.000
4	13.942	0.003
5	3.580	0.083

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.163	0.013	1	0.299	0.596
6	POPULATN	-0.376	0.005	1	1.813	0.205
7	TIME	0.830	0.002	1	24.314	0.000

Information Criteria

AIC : 248.317
 AIC (Corrected) : 254.317
 Schwarz's BIC : 252.180

Step Number : 4
 R : 0.998
 R-square : 0.995
 Term Entered : TIME

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	-0.040	0.016	-1.137	0.002	1
4	UNEMPLOY	-2.088	0.290	-0.556	0.071	1
5	ARMFORCE	-1.015	0.184	-0.201	0.318	1
7	TIME	1887.410	382.766	2.559	0.002	1

In	F-ratio	p-value
1		
3	5.953	0.033
4	51.870	0.000
5	30.496	0.000
7	24.314	0.000

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.143	0.013	1	0.208	0.658
6	POPULATN	-0.150	0.004	1	0.230	0.642

Information Criteria

AIC : 231.655
 AIC (Corrected) : 240.988
 Schwarz's BIC : 236.291

Dependent Variable : TOTAL
 N : 16
 Multiple R : 0.998
 Squared Multiple R : 0.995
 Adjusted Squared Multiple R : 0.994
 Standard Error of Estimate : 279.396

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	-3.599E+006	740632.644	0.000	.	-4.859
GNP	-0.040	0.016	-1.137	0.002	-2.440
UNEMPLOY	-2.088	0.290	-0.556	0.071	-7.202
ARMFORCE	-1.015	0.184	-0.201	0.318	-5.522
TIME	1887.410	382.766	2.559	0.002	4.931

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.001
GNP	0.033
UNEMPLOY	0.000
ARMFORCE	0.000
TIME	0.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	1.842E+008	4	4.604E+007	589.757	0.000
Residual	858680.406	11	78061.855		

The steps proceed as follows:

- At step 0, no variables are in the model. *GNP* has the largest simple correlation and *F-ratio*, so SYSTAT enters it at step 1. Note at this step that the partial correlation, *Part. Corr.*, is the simple correlation of each predictor with *TOTAL*.
- With *GNP* in the equation, *UNEMPLOY* is now the best candidate.
- The *F-ratio* for *ARMFORCE* is 3.58 when *GNP* and *UNEMPLOY* are included in the model.
- SYSTAT finishes by entering *TIME*.

In four steps, SYSTAT entered four predictors. None was removed, resulting in a final equation with a constant and four predictors. For this final model, SYSTAT uses all cases with complete data for *GNP*, *UNEMPLOY*, *ARMFORCE*, and *TIME*. Thus, when some values in the sample are missing, the sample size may be larger here than for the last step in the stepwise process (there, cases are omitted if any value is missing among the six candidate variables). If you do not want to stop here, you could move more variables in (or out) using interactive stepping.

AIC, *AIC (Corrected)* and Schwarz's *BIC* values of the final model are less than the corresponding information criteria values of the models fit in previous steps. Thus the final model is a better approximation of the true model in comparison to the models in previous steps.

Example 6

Interactive Stepwise Regression

Interactive stepping helps you to explore model building in more detail. With data that are as highly intercorrelated as the *LONGLEY* data, interactive stepping reveals the dangers of thinking that the automated result is the only acceptable subset model. In

this example, we use interactive stepping to explore the *LONGLEY* data further. That is, after specifying a model that includes all of the candidate variables available, we request backward stepping by selecting Stepwise, Backward, and Interactive in the Regression Estimation tab. After reviewing the results at each step, we use Step to move a variable in (or out) of the model. When finished, we select Stop for the final model.

The input is:

```
REGRESS
USE LONGLEY
MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
              ARMFORCE + POPULATN + TIME
START / BACK
```

The output is:

Stepwise Selection of Variables

```
Step Number : 0
R           : 0.998
R-square    : 0.995
```

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
2	DEFLATOR	15.062	84.915	0.046	0.007	1
3	GNP	-0.036	0.033	-1.014	0.001	1
4	UNEMPLOY	-2.020	0.488	-0.538	0.030	1
5	ARMFORCE	-1.033	0.214	-0.205	0.279	1
6	POPULATN	-0.051	0.226	-0.101	0.003	1
7	TIME	1829.151	455.478	2.480	0.001	1

In	F-ratio	p-value
1		
2	0.031	0.863
3	1.144	0.313
4	17.110	0.003
5	23.252	0.001
6	0.051	0.826
7	16.127	0.003

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
	none					

Information Criteria

```
AIC           | 235.235
AIC (Corrected) | 255.806
Schwarz's BIC  | 241.416
```

We begin with all variables in the model. We remove *DEFLATOR* because it has an unusually low tolerance and *F-ratio* value.

The input is:

STEP DEFLATOR

The output is:

Dependent Variable : TOTAL
 Minimum Tolerance for Entry into Model : 0.000
 Backward Stepwise with Alpha-to-Enter : 0.150
 Backward Stepwise with Alpha-to-Remove : 0.150

Step Number : 1
 R : 0.998
 R-square : 0.995
 Term Removed : DEFLATOR

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant		0.024	-0.905	0.001	1
3	GNP	-0.032	0.386	-0.525	0.043	1
4	UNEMPLOY	-1.972	0.191	-0.202	0.317	1
5	ARMFORCE	-1.020	0.162	-0.154	0.004	1
6	POPULATN	-0.078	425.283	2.459	0.001	1
7	TIME	1814.101				

In	F-ratio	p-value
1		
3	1.744	0.216
4	26.090	0.000
5	28.564	0.000
6	0.230	0.642
7	18.196	0.002

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.059	0.007	1	0.031	0.863

Information Criteria

AIC : 233.291
 AIC (Corrected) : 247.291
 Schwarz's BIC : 238.699

POPULATN has the lowest *F-ratio* and, again, a low tolerance.

The input is:

STEP POPULATN

The output is:

Step Number : 2
 R : 0.998
 R-square : 0.995
 Term Removed : POPULATN

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	-0.040	0.016	-1.137	0.002	1
4	UNEMPLOY	-2.088	0.290	-0.556	0.071	1
5	ARMFORCE	-1.015	0.184	-0.201	0.318	1
7	TIME	1887.410	382.766	2.559	0.002	1

In	F-ratio	p-value
1		
3	5.953	0.033
4	51.870	0.000
5	30.496	0.000
7	24.314	0.000

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.143	0.013	1	0.208	0.658
6	POPULATN	-0.150	0.004	1	0.230	0.642

Information Criteria

AIC : 231.655
 AIC (Corrected) : 240.988
 Schwarz's BIC : 236.291

GNP and TIME both have low tolerance values. They could be highly correlated with one another, so we will take each out and examine the behavior of the other when we do.

The input is:

STEP TIME
 STEP TIME
 STEP GNP

The output is:

Step Number : 3
 R : 0.993
 R-square : 0.985
 Term Removed : TIME

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	0.041	0.002	1.154	0.318	1
4	UNEMPLOY	-0.797	0.213	-0.212	0.385	1
5	ARMFORCE	-0.483	0.255	-0.096	0.486	1

Linear Models I: Linear Regression

In	F-ratio	p-value
1		
3	341.684	0.000
4	13.942	0.003
5	3.580	0.083

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.163	0.013	1	0.299	0.596
6	POPULATN	-0.376	0.005	1	1.813	0.205
7	TIME	0.830	0.002	1	24.314	0.000

Information Criteria

AIC : 248.317
 AIC (Corrected) : 254.317
 Schwarz's BIC : 252.180

Step Number : 4
 R : 0.998
 R-square : 0.995
 Term Entered : TIME

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
3	GNP	-0.040	0.016	-1.137	0.002	1
4	UNEMPLOY	-2.088	0.290	-0.556	0.071	1
5	ARMFORCE	-1.015	0.184	-0.201	0.318	1
7	TIME	1887.410	382.766	2.559	0.002	1

In	F-ratio	p-value
1		
3	5.953	0.033
4	51.870	0.000
5	30.496	0.000
7	24.314	0.000

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	0.143	0.013	1	0.208	0.658
6	POPULATN	-0.150	0.004	1	0.230	0.642

Information Criteria

AIC : 231.655
 AIC (Corrected) : 240.988
 Schwarz's BIC : 236.291

Step Number : 5
 R : 0.996
 R-square : 0.993
 Term Removed : GNP

In	Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	df
1	Constant					
4	UNEMPLOY	-1.470	0.167	-0.391	0.301	1
5	ARMFORCE	-0.772	0.184	-0.153	0.450	1
7	TIME	956.380	35.525	1.297	0.257	1

In	F-ratio	p-value
1		
4	77.320	0.000
5	17.671	0.001
7	724.765	0.000

Out	Effect	Partial Correlation	Tolerance	df	F-ratio	p-value
2	DEFLATOR	-0.031	0.014	1	0.011	0.920
3	GNP	-0.593	0.002	1	5.953	0.033
6	POPULATN	-0.505	0.009	1	3.768	0.078

Information Criteria

AIC	236.576
AIC (Corrected)	242.576
Schwarz's BIC	240.439

We are comfortable with the tolerance values in both models with three variables. With *TIME* in the model, the smallest *F-ratio* is 17.671, and with *GNP* in the model, the smallest *F-ratio* is 3.580. Furthermore, with *TIME*, the squared multiple correlation is 0.993, and with *GNP*, it is 0.985. Let's stop the stepping and view more information about the last model.

The input is:

STOP

The output is:

Dependent Variable	TOTAL
N	16
Multiple R	0.996
Squared Multiple R	0.993
Adjusted Squared Multiple R	0.991
Standard Error of Estimate	332.084

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	-1.797E+006	68641.553	0.000		-26.183
UNEMPLOY	-1.470	0.167	-0.391	0.301	-8.793
ARMFORCE	-0.772	0.184	-0.153	0.450	-4.204
TIME	956.380	35.525	1.297	0.257	26.921

Regression Coefficients B = $(X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.000
UNEMPLOY	0.000
ARMFORCE	0.001
TIME	0.000

Confidence Interval for Regression Coefficients

95.0% Confidence Interval

Effect	Coefficient	Lower	Upper	VIF
CONSTANT	-1.797E+006	-1.947E+006	-1.648E+006	
UNEMPLOY	-1.470	-1.834	-1.106	3.318
ARMFORCE	-0.772	-1.173	-0.372	2.223
TIME	956.380	878.978	1033.782	3.891

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	1.837E+008	3	6.123E+007	555.209	0.000
Residual	1.323E+006	12	110280.062		

Our final model includes only *UNEMPLOY*, *ARMFORCE*, and *TIME*. Notice that its multiple correlation (0.996) is not significantly smaller than that for the automated stepping (0.998).

The input is:

```

REGRESS
USE LONGLEY
MODEL TOTAL=CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
                    ARMFORCE + POPULATN + TIME

START / BACK
STEP DEFLATOR
STEP POPULATN
STEP TIME
STEP TIME
STEP GNP
STOP

```

Example 7

Testing whether a Single Coefficient Equals Zero

Most regression programs print tests of significance for each coefficient in an equation. SYSTAT has a powerful additional feature—post hoc tests of regression coefficients. To demonstrate these tests, we use the *LONGLEY* data and examine whether the *DEFLATOR* coefficient differs significantly from 0.

The input is:

```

REGRESS
USE LONGLEY
MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
              ARMFORCE + POPULATN + TIME
ESTIMATE / TOL=.00001
HYPOTHESIS
EFFECT DEFLATOR
TEST

```

The output is:

```

Dependent Variable      TOTAL
N                        16
Multiple R              0.998
Squared Multiple R      0.995
Adjusted Squared Multiple R 0.992
Standard Error of Estimate 304.854

```

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	-3.482E+006	890420.384	0.000	.	-3.911
DEFLATOR	15.062	84.915	0.046	0.007	0.177
GNP	-0.036	0.033	-1.014	0.001	-1.070
UNEMPLOY	-2.020	0.488	-0.538	0.030	-4.136
ARMFORCE	-1.033	0.214	-0.205	0.279	-4.822
POPULATN	-0.051	0.226	-0.101	0.003	-0.226
TIME	1829.151	455.478	2.480	0.001	4.016

Regression Coefficients B = $(X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.004
DEFLATOR	0.863
GNP	0.313
UNEMPLOY	0.003
ARMFORCE	0.001
POPULATN	0.826
TIME	0.003

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	1.842E+008	6	3.070E+007	330.285	0.000
Residual	836424.056	9	92936.006		

Test for effect called: DEFLATOR

Contrast Estimate

Hypothesis	Estimate (AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	15.062	84.915	12.939	17.185

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	2923.976	1	2923.976	0.031	0.863
Error	836424.056	9	92936.006		

Notice that the error sum of squares (836424.056) is the same as the output residual sum of squares at the bottom of the ANOVA table. The probability level (0.863) is the same also. This probability level (> 0.05) indicates that the regression coefficient for *DEFLATOR* does not differ from 0.

You can test all of the coefficients in the equation this way, individually, or choose All to generate separate hypothesis tests for each predictor or type:

```
HYPOTHESIS
ALL
TEST
```

Example 8**Testing whether Multiple Coefficients Equal Zero**

You may wonder why you need to bother with testing when the regression output gives you hypothesis test results.

The input is:

```
REGRESS
USE LONGLEY
MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
               ARMFORCE + POPULATN + TIME
ESTIMATE / TOL=.00001
HYPOTHESIS
EFFECT DEFLATOR & GNP
TEST
```

The output is:

Test for effect called: DEFLATOR and GNP

A Matrix

	1	2	3	4	5
1	0.000	1.000	0.000	0.000	0.000
2	0.000	0.000	1.000	0.000	0.000

A Matrix

	6	7
1	0.000	0.000
2	0.000	0.000

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A1	15.062	84.915	12.939	17.185
A2	-0.036	0.033	-0.037	-0.035

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	2923.976	1	2923.976	0.031	0.863
A2	106306.259	1	106306.259	1.144	0.313
A	149295.592	2	74647.796	0.803	0.478
Error	836424.056	9	92936.006		

Here, the error sum of squares is the same as that for the model, but the hypothesis sum of squares is different. We just tested the hypothesis that the *DEFLATOR* and *GNP* coefficients simultaneously are 0.

The **A** matrix printed above the test specifies the hypothesis that we tested. It has two degrees of freedom (see the *F-ratio*) because the **A** matrix has two rows—one for each coefficient. If you know some matrix algebra, you can see that the matrix product **AB** using this **A** matrix and **B** as a column matrix of regression coefficients picks up only two coefficients: *DEFLATOR* and *GNP*. Notice that our hypothesis had the following matrix equation: $\mathbf{AB} = \mathbf{0}$, where **0** is a null matrix.

If you don't know matrix algebra, don't worry; the ampersand method is equivalent. You can ignore the **A** matrix in the output.

Two Coefficients with an A Matrix

If you are experienced with matrix algebra, however, you can specify your own matrix by using **AMATRIX**. When typing the matrix, be sure to separate cells with spaces and press Enter between rows. The following simultaneously tests that *DEFLATOR* = 0 and *GNP* = 0:

```

HYPOTHESIS
AMATRIX [0 1 0 0 0 0 0;
          0 0 1 0 0 0 0]
TEST

```

You get the same output as above.

Why bother with **AMATRIX** when you can use **EFFECT**? Because in the **A** matrix, you can use any numbers, not just 0's and 1's. Here is a bizarre matrix:

```
1.0 3.0 0.5 64.3 3.0 2.0 0.0
```

You may not want to test this kind of hypothesis on the *LONGLEY* data, but there are important applications in the analysis of variance where you might.

Example 9

Testing Nonzero Null Hypotheses

You can test nonzero null hypotheses with a **D** matrix, often in combination using **CONTRAST** or **AMATRIX**. Here, we test whether the *DEFLATOR* coefficient significantly differs from 30.

The input is:

```
REGRESS
USE LONGLEY
MODEL TOTAL = CONSTANT + DEFLATOR + GNP + UNEMPLOY +,
              ARMFORCE + POPULATN + TIME
ESTIMATE / TOL=.00001
HYPOTHESIS
AMATRIX [0 1 0 0 0 0 0]
DMATRIX [30]
TEST
```

The output is:

A Matrix

1	2	3	4	5
0.000	1.000	0.000	0.000	0.000

A Matrix

6	7
0.000	0.000

Null Hypothesis Value for D

30.000

Contrast Estimate

Hypothesis	Estimate (AB-D)	Standard Error	95.0% Confidence Interval Lower	Upper
A	-14.938	84.915	-17.061	-12.815

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	2876.128	1	2876.128	0.031	0.864
Error	836424.056	9	92936.006		

The commands that test whether *DEFLATOR* differs from 30 can be performed more efficiently using SPECIFY:

```
HYPOTHESIS
  SPECIFY DEFLATOR=30
  TEST
```

Example 10

Regression with Ecological or Grouped Data

If you have aggregated data, weight the regression by a count variable. This variable should represent the counts of observations (n) contributing to the i th case. If n is not an integer, SYSTAT truncates it to an integer before using it as a weight. The regression results are identical to those produced if you had typed in each case.

We use, for this example, an ecological or grouped data file, *PLANTS*.

The input is:

```
REGRESS
  USE PLANTS
  FREQ COUNT
  MODEL CO2 = CONSTANT + SPECIES
  ESTIMATE
```

The output is:

```
Dependent Variable : CO2
N : 76
Multiple R : 0.757
Squared Multiple R : 0.573
Adjusted Squared Multiple R : 0.567
Standard Error of Estimate : 0.729
```

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
CONSTANT	13.738	0.204	0.000	.	67.273
SPECIES	-0.466	0.047	-0.757	1.000	-9.961

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.000
SPECIES	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	13.738	13.331	14.144	.
SPECIES	-0.466	-0.559	-0.372	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	52.660	1	52.660	99.223	0.000
Residual	39.274	74	0.531		

Example 11**Regression without the Constant**

To regress without the constant (intercept) term, or through the origin, remove the constant from the list of independent variables, REGRESS adjusts accordingly.

The input is:

```
REGRESS
USE LONGLEY
MODEL TOTAL = DEFLATOR+ GNP+ UNEMPLOY+ ARMFORCE+ POPULATN +
              TIME
ESTIMATE
```

The output is:

```
Model Contains no Constant
Dependent Variable : TOTAL
N : 16
Multiple R : 1.000
Squared Multiple R : 1.000
Adjusted Squared Multiple R : 1.000
Standard Error of Estimate : 475.166
```

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
DEFLATOR	-52.994	129.545	-0.083	0.000	-0.409
GNP	0.071	0.030	0.434	0.000	2.356
UNEMPLOY	-0.423	0.418	-0.021	0.007	-1.014
ARMFORCE	-0.573	0.279	-0.024	0.025	-2.052
POPULATN	-0.414	0.321	-0.745	0.000	-1.289
TIME	48.418	17.689	1.447	0.000	2.737

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
DEFLATOR	0.691
GNP	0.040
UNEMPLOY	0.335
ARMFORCE	0.067
POPULATN	0.226
TIME	0.021

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
DEFLATOR	-52.994	-341.638	235.650	12425.514
GNP	0.071	0.004	0.138	10290.435
UNEMPLOY	-0.423	-1.354	0.507	136.224
ARMFORCE	-0.573	-1.194	0.049	39.983
POPULATN	-0.414	-1.130	0.302	101193.162
TIME	48.418	9.003	87.832	84709.950

Correlation Matrix of Regression Coefficients

	DEFLATOR	GNP	UNEMPLOY	ARMFORCE	POPULATN
DEFLATOR	1.000				
GNP	-0.852	1.000			
UNEMPLOY	-0.714	0.830	1.000		
ARMFORCE	-0.289	0.041	0.347	1.000	
POPULATN	0.644	-0.945	-0.829	0.048	1.000
TIME	-0.762	0.985	0.850	0.009	-0.986

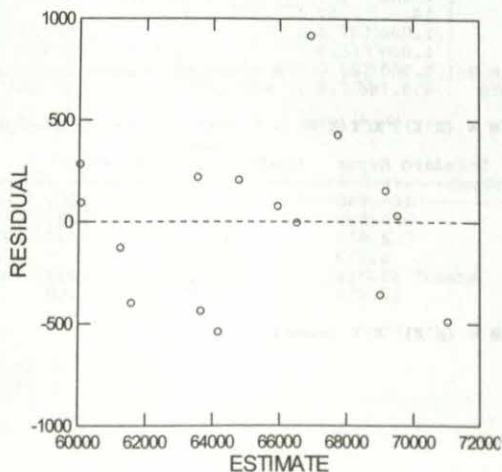
Correlation Matrix of Regression Coefficients

	TIME
TIME	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	6.844E+010	6	1.141E+010	50523.396	0.000
Residual	2.258E+006	10	225782.260		

Plot of Residuals vs Predicted Values



Some users are puzzled when they see a model without a constant having a higher multiple correlation than a model that includes a constant. How can a regression with fewer parameters predict "better" than another? It doesn't. The total sum of squares must be redefined for a regression model with zero intercept. It is no longer centered about the mean of the dependent variable. Other definitions of sum of squares can lead to strange results, such as negative multiple correlations. If your constant is actually near 0, then including or excluding the constant makes little difference in the output. Kvålseth (1985) discusses the issues involved in summary statistics for zero-intercept regression models. The definition used in SYSTAT is Kvålseth's formula 7. This was chosen because it retains its PRE (percentage reduction of error) interpretation and is guaranteed to be in the (0,1) interval.

How, then, do you test the significance of a constant in a regression model? Include a constant in the model as usual and look at its test of significance.

If you have a zero-intercept model where it is appropriate to compute a coefficient of determination and other summary statistics about the centered data, use General Linear Model and select Mixture model. This option provides Kvålseth's formula 1 for R^2 and uses centered total sum of squares for other summary statistics.

Example 12

Regression using SSCP, Covariance or Correlation matrices

You can regress the data which is in the form of correlation, covariance and SSCP matrices by directly inputting the matrix in SYSTAT. Along with it specify the number of cases and also dependent and independent variables. The data set used in this example is a covariance matrix of *NFL* data. In this example, we build a multiple regression model to predict dependent variable *RATING* using five independent variables.

We compute a covariance matrix, save it and use it in the regression analysis:

The input is:

```
CORR
USE NFL
SAVE NFLCOV
COVARIANCE ATTEMPTS COMPLETIONS YARDS TDS INTS RATING
REGRESS
USE NFLCOV
MODEL RATING = ATTEMPTS+COMPLETIONS+YARDS+TDS+INTS /N=21
ESTIMATE
```


The output is:

```

Dependent Variable      : RATING
N                        : 21
Multiple R              : 0.977
Squared Multiple R      : 0.955
Adjusted Squared Multiple R : 0.940
Standard Error of Estimate : 0.940

```

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Coefficient	Tolerance	t
ATTEMPTS	-0.021	0.002	-7.121	0.005	-9.593
COMPLETIONS	0.020	0.004	4.139	0.006	5.677
YARDS	0.001	0.000	2.952	0.008	4.725
TDS	0.060	0.012	1.088	0.064	5.027
INTS	-0.079	0.013	-0.978	0.115	-6.044

Regression Coefficients B = $(X'X)^{-1}X'Y$ (contd...)

Effect	p-value
ATTEMPTS	0.000
COMPLETIONS	0.000
YARDS	0.000
TDS	0.000
INTS	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
ATTEMPTS	-0.021	-0.025	-0.016	183.065
COMPLETIONS	0.020	0.012	0.027	176.600
YARDS	0.001	0.001	0.002	129.628
TDS	0.060	0.035	0.086	15.560
INTS	-0.079	-0.106	-0.051	8.695

Correlation Matrix of Regression Coefficients

	ATTEMPTS	COMPLETIONS	YARDS	TDS	INTS
ATTEMPTS	1.000				
COMPLETIONS	-0.776	1.000			
YARDS	-0.280	-0.362	1.000		
TDS	0.352	-0.083	-0.570	1.000	
INTS	-0.530	0.670	-0.259	-0.343	1.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	280.002	5	56.000	63.444	0.000
Residual	13.240	15	0.883		

In case of correlation matrix, the raw and standardized coefficients are the same. The Include constant option is disabled because the matrices are already centered. SYSTAT requires the original sample size in order to compute the degrees of freedom.

If we use the following input the two outputs will have identical results except residuals. The coefficients are the same; it takes degrees of freedom from the Cases

which is the original sample size. The standardized coefficients (*Std. Coefficient*) for model with constant are the same for covariance and SSCP matrices.

```
USE NFL
REGRESS
MODEL RATING= CONSTANT+ATTEMPTS+COMPLETIONS+YARDS+TDS+
INTS
ESTIMATE
```

Example 13 Seemingly Unrelated Regression Equations

This example taken from Judge, et al. (1988), illustrates the efficiency of combining two linear models which are contemporaneously correlated, over the individual models themselves. The two models and the combined model are given below. The SYSTAT data file *JUDGEHILL*, used in the example is obtained on appending data for the two models. It contains two indicator variables X_{11} and X_{21} representing the cases obtained from the first and second models respectively. X_{12} and X_{22} represent the market values of a certain product of two different companies with capital stocks X_{13} and X_{23} respectively. The dependent variable Y represents the investment figures for the two companies. The data set is fictitious.

The individual models are

$$y_1 = x_{11}\beta_{11} + x_{12}\beta_{12} + x_{13}\beta_{13} \quad (1)$$

$$y_2 = x_{21}\beta_{21} + x_{22}\beta_{22} + x_{23}\beta_{23} \quad (2)$$

The combined model is

$$y = x_{11}\beta_{11} + x_{12}\beta_{12} + x_{13}\beta_{13} + x_{21}\beta_{21} + x_{22}\beta_{22} + x_{23}\beta_{23} \quad (3)$$

Since the individual models are correlated, the errors in the combined model are correlated. In order to carry out simple linear regression, we should make transformations on the data such that the errors become uncorrelated. We first estimate the covariance matrix of the combined model and use it for transformation. The covariance matrix of the combined model can be derived as the Kronecker product of the covariance matrix between the errors of the individual models and the identity matrix of order equal to the number of observations in the individual models. An estimate of covariance matrix between the errors of the individual models can be

obtained using the sample covariance matrix between the residuals of the fitted individual models. An adjustment for this matrix is done, since SYSTAT computes covariance matrix with a number one less than the number of observations used, in the denominator. But the degrees of freedom of the residual sum of squares should be used in the denominator for the estimation of covariance matrix. Now use this estimate in computing an estimate of the covariance matrix of the combined model. The transformed data with uncorrelated errors is obtained by multiplying the inverse of the square root matrix of the estimated covariance matrix with the original data. For details on the covariance structure of the errors in the combined model, refer Judge, et al. (1988).

The input is:

```
USE JUDGEHILL
SELECT (case < 21)
REGRESS
MODEL Y = X11+X12+X13
SAVE 1.SYZ / RESIDUALS
ESTIMATE
SELECT (case > 20)
REGRESS
MODEL Y = X21+X22+X23
SAVE 2.SYZ / RESIDUALS
ESTIMATE
MERGE 1.SYZ (RESIDUAL) 2.SYZ (RESIDUAL)
DSAVE RESID
CORR
SAVE COVAR12
COVARIANCE RESIDUAL__1 RESIDUAL__2
USE COVAR12 / MAT = SIG MTYPE = NUMERIC
MAT SIG = SIG( ; RESIDUAL__1 RESIDUAL__2 )
MAT SIG = SIG*19/17
MAT SIG = FOLD(SIG)
MAT SIG1 = KRON(SIG,I(20))
MAT SIGINV = INV(CHOL(SIG1))
USE JUDGEHILL / MAT = DATA MTYPE = NUMERIC
MAT DATA = DATA( ; X11 X12 X13 X21 X22 X23 Y )
MAT TRANS_DATA = SIGINV*DATA
MSAVE TRANS_DATA
USE TRANS_DATA
REGRESS
MODEL Y = X11+X12+X13+X21+X22+X23
ESTIMATE
```


The output is:

OLS Regression

Data for the following results were selected according to
SELECT (case < 21)

Model Contains no Constant

Dependent Variable	Y
N	20
Multiple R	0.976
Squared Multiple R	0.952
Adjusted Squared Multiple R	0.946
Standard Error of Estimate	29.324

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
X11	5.279	32.997	0.043	0.039	0.160
X12	0.024	0.016	0.379	0.041	1.441
X13	0.156	0.027	0.593	0.268	5.777

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
X11	0.875
X12	0.168
X13	0.000

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	290617.309	3	96872.436	112.652	0.000
Residual	14618.730	17	859.925		

Durbin-Watson D Statistic	1.496
First Order Autocorrelation	0.238

Information Criteria

AIC	196.644
AIC (Corrected)	199.311
Schwarz's BIC	200.627

Residuals have been saved.

OLS Regression

Data for the following results were selected according to
SELECT (case > 20)

Model Contains no Constant

Dependent Variable	Y
N	20
Multiple R	0.970
Squared Multiple R	0.941
Adjusted Squared Multiple R	0.934
Standard Error of Estimate	13.652

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
X21	-0.550	10.718	-0.011	0.081	-0.051
X22	0.062	0.021	0.845	0.043	2.964
X23	0.073	0.075	0.147	0.151	0.970

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
X21	0.960
X22	0.009
X23	0.346

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	50730.973	3	16910.324	90.729	0.000
Residual	3168.523	17	186.384		

Durbin-Watson D Statistic : 2.107
First Order Autocorrelation : -0.066

Information Criteria

AIC : 166.063
AIC (Corrected) : 168.730
Schwarz's BIC : 170.046

Residuals have been saved.

OLS Regression
Model Contains no Constant

Dependent Variable : Y
N : 40
Multiple R : 0.956
Squared Multiple R : 0.915
Adjusted Squared Multiple R : 0.902
Standard Error of Estimate : 1.003

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance	t
X11	2.694	30.325	0.032	0.019	0.089
X12	0.024	0.015	0.568	0.020	1.616
X13	0.161	0.026	0.897	0.118	6.156
X21	-3.234	9.968	-0.083	0.039	-0.324
X22	0.065	0.019	1.173	0.021	3.409
X23	0.081	0.070	0.216	0.071	1.151

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
X11	0.930
X12	0.115
X13	0.000
X21	0.748
X22	0.002
X23	0.258

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	366.960	6	61.160	60.799	0.000
Residual	34.202	34	1.006		

Durbin-Watson D Statistic : 2.133
First Order Autocorrelation : -0.082

Information Criteria

AIC : 121.251
AIC (Corrected) : 124.751
Schwarz's BIC : 133.073

The output displayed contains the analysis of three simple linear regressions carried over the models (1), (2) and the transformation of the model (3) respectively. The estimate of σ^2 is given by the Residual Mean Squares. For the first model the estimate is 859.925 and for the second model it is 186.384. Similarly for the transformed model the estimate is 1.006, which is much better than those of the first two models.

Example 14

Prediction of New Observations

SYSTAT predicts values of the dependent variable for given values of the independent variables, along with its standard error, upper and lower confidence and prediction limits. You can input the given values through a .syz file. The names of the variables should be the same in both original and new files. Do not input values for the dependent (response) variable. In the data file *NFL* we predict the response variable *RATING* for new values of *ATTEMPTS* *COMPLETIONS* *YARDS* *TDS* and *INTS*. The new values for prediction are in file *NEWNFL*.

The input is:

```
REGRESS
USE NFL
MODEL RATING = CONSTANT+ATTEMPTS+COMPLETIONS+YARDS+TDS+,
INTS
ESTIMATE
PREDICT NEWNFL
```

The output is:

Dependent Variable	RATING
N	21
Multiple R	0.977
Squared Multiple R	0.955
Adjusted Squared Multiple R	0.940
Standard Error of Estimate	1.050

Regression Coefficients $B = (X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Coefficient	Std. Tolerance	t
CONSTANT	84.346	0.710	0.000	.	118.823
ATTEMPTS	-0.021	0.002	-7.121	0.005	-9.593
COMPLETIONS	0.020	0.004	4.139	0.006	5.677
YARDS	0.001	0.000	2.952	0.008	4.725
TDS	0.060	0.012	1.088	0.064	5.027
INTS	-0.079	0.013	-0.978	0.115	-6.044

Regression Coefficients $B = (X'X)^{-1}X'Y$ (contd...)

Effect	p-value
CONSTANT	0.000
ATTEMPTS	0.000
COMPLETIONS	0.000
YARDS	0.000
TDS	0.000
INTS	0.000

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
CONSTANT	84.346	82.833	85.859	.
ATTEMPTS	-0.021	-0.025	-0.016	183.065
COMPLETIONS	0.020	0.012	0.027	176.600
YARDS	0.001	0.001	0.002	129.628
TDS	0.060	0.035	0.086	15.560
INTS	-0.079	-0.106	-0.051	8.695

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	350.002	5	70.000	63.444	0.000
Residual	16.550	15	1.103		

New Values

ATTEMPTS	COMPLETIONS	YARDS	TDS	INTS
5100.000	4122.000	54213.000	250.000	201.000
2333.000	2431.000	26754.000	231.000	198.000
4532.000	1342.000	65742.000	145.000	114.000
2234.000	1675.000	54897.000	121.000	176.000
5467.000	3421.000	15478.000	117.000	123.000
3249.000	5643.000	38765.000	187.000	127.000
4167.000	2318.000	18762.000	327.000	149.000

Prediction of New Values

Predicted	Standard Error	95.0% Confidence Interval		95.0% Prediction Interval
		Lower	Upper	Lower

123.387	4.674	113.425	133.350	113.176
114.011	4.129	105.209	122.813	104.929
93.926	11.292	69.857	117.994	69.753
129.114	8.880	110.186	148.041	110.054
54.624	5.301	43.326	65.923	43.106
175.881	12.266	149.737	202.025	149.641
74.048	3.899	65.739	82.358	65.442

Prediction of New Values (contd...)

Upper
133.598
123.093
118.098
148.173
66.142
202.121
82.654

Confidence limits are limits for a mean response at a level of predictor values, whereas *prediction limits* are limits for the response of a randomly selected unit from the population at a certain level of predictor values.

You can also save the new predicted values along with the new set of values of independent variables in the model.

The input is:

```

REGRESS
USE NFL
MODEL RATING= CONSTANT+ATTEMPTS+COMPLETIONS+YARDS+TDS+ INTS
ESTIMATE
SAVE PREDICT/PREDICT,NEWDATA
PREDICT NEWNFL

```

Example 15

Ridge Regression Analysis

In this example, we build a multiple regression model to predict dependent variable *TOTAL* using values of six independent variables - *DEFLATOR*, *GNP*, *UNEMPLOY*, *ARMFORCE*, *POPULATN*, *TIME*. The data were originally used by Longley (1967) to test the robustness of least-squares packages to multicollinearity and other sources of ill-conditioning.

The input is:

```

RIDGE REG
USE RLONGLEY
MODEL TOTAL = DEFLATOR+GNP+UNEMPLOY+ARMFORCE+POPULATN+TIME
ESTIMATE/ LMIN=.2 LMAX=.6 LSTEP=.1

```

The output is:

```

Hoerl-Kennard-Baldwin (HKB) Estimator      : 0.000
Lawless & Wang (LW) Estimator              : 0.003
Minimum Value of Generalized Cross Validation (GCV) is at Lambda : 0.600

```

Standardized Ridge Coefficients

LAMBDA	DEFLATOR	GNP	UNEMPLOY	ARMFORCE	POPULATN	TIME
0.200	0.241	0.276	-0.115	0.011	0.230	0.251
0.300	0.229	0.257	-0.075	0.034	0.221	0.234
0.400	0.220	0.243	-0.048	0.048	0.213	0.223
0.500	0.213	0.232	-0.029	0.057	0.206	0.214
0.600	0.206	0.223	-0.015	0.063	0.200	0.207

Unstandardized Ridge Coefficients

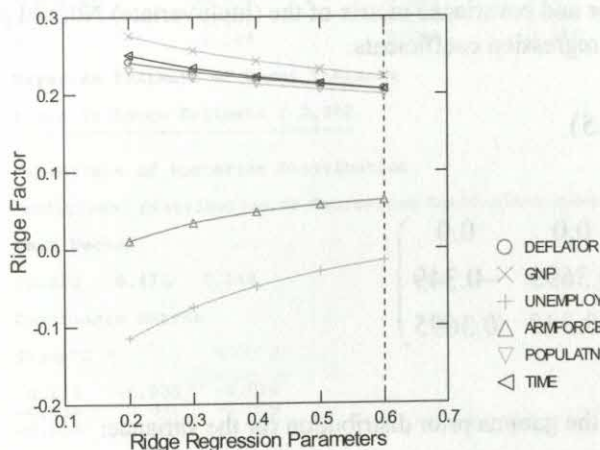
LAMBDA	CONSTANT	DEFLATOR	GNP	UNEMPLOY	ARMFORCE	POPULATN	TIME
0.200	-320.392	0.079	0.010	-0.004	0.001	0.116	0.185
0.300	-295.796	0.075	0.009	-0.003	0.002	0.112	0.173
0.400	-278.868	0.072	0.009	-0.002	0.002	0.108	0.164
0.500	-265.733	0.069	0.008	-0.001	0.003	0.104	0.158
0.600	-254.844	0.067	0.008	-0.001	0.003	0.101	0.152

Estimate of Bias Vector for Standardized Coefficients

LAMBDA	DEFLATOR	GNP	UNEMPLOY	ARMFORCE	POPULATN	TIME
0.200	-0.027	-0.045	0.100	0.061	-0.019	-0.043
0.300	-0.031	-0.048	0.096	0.053	-0.026	-0.040
0.400	-0.033	-0.049	0.089	0.044	-0.030	-0.039
0.500	-0.035	-0.049	0.081	0.035	-0.032	-0.039
0.600	-0.037	-0.050	0.073	0.028	-0.034	-0.040

Ridge regression estimators have a bias, but a smaller mean square error than that of ordinary least-squares estimates. SYSTAT produces estimates of the bias vector for all lambdas and covariance matrix of standardized ridge regression coefficients for a given lambda or the first value from a set of lambda values.

Ridge Regression Parameters



Example 16

Bayesian Regression

To illustrate the Bayesian regression, we use the data related to the Cobb-Douglas production function (Judge et al., 1988).

The Cobb-Douglas production function is given by:

$$Q = \alpha L^{\beta_1} K^{\beta_2} \exp(\varepsilon)$$

where Q, L, and K represent the output, 'labor', and 'capital invested' respectively. When a logarithmic transformation is used for α , L, and K, we obtain the linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where $Y = \log(Q)$, $\beta_0 = \log(\alpha)$, $X_1 = \log(L)$, and $X_2 = \log(K)$.

The data set consists of 20 observations and the purpose here is to study the effect of labor and capital on the output Y. To fit a Bayesian regression model to this data, we have to specify the parameters of the prior distribution.

The mean vector and covariance matrix of the (multivariate) Normal prior distribution of the regression coefficients:

$$b_0 = (5.0, 0.5, 0.5)$$

and

$$V_0 = \begin{pmatrix} 924.0 & 0.0 & 0.0 \\ 0 & 0.3695 & -0.349 \\ 0 & -0.349 & 0.3695 \end{pmatrix}$$

The parameters of the gamma prior distribution for the variance:

Scale = 4.0 and Shape = 0.0754.

The input is:

```
BAYESIAN
USE COBDOUG
MODEL Y=CONSTANT +X1+X2
SAVE BAYOUT1 / COEFFICIENTS
ESTIMATE / MEAN = [5;0.5;0.5] VAR = [924 0 0; 0 0.3695,
-0.349; 0 -0.349 0.3695] SCALE = 4 SHAPE = 0.0754
```

The output is:

Normal Prior Mean

```
5.000  0.500  0.500
```

Normal Prior Covariance Matrix

```
924.000  0.000  0.000
0.000  0.370 -0.349
0.000 -0.349  0.370
```

Gamma Prior Parameters

```
Scale Parameter  4.000
Shape Parameter  0.075
```

Bayesian Estimate of Regression Coefficients and Credible Intervals

Effect	Coefficient	Standard Error	95% Credible Interval	
			Lower	Upper
CONSTANT	10.028	0.101	9.830	10.227
X1	0.476	0.077	0.325	0.627
X2	0.548	0.083	0.384	0.711

Bayesian Estimate of Error Variance

Error Variance Estimate : 0.082

Parameters of Posterior Distribution

Conditional Distribution of Regression Coefficient given Sigma follows Multivariate Normal with

Mean Vector

10.028 0.476 0.548

Covariance Matrix

Sigma^2 *

0.123	-0.035	-0.014
-0.035	0.071	-0.059
-0.014	-0.059	0.083

Marginal Distribution of Regression Coefficient follows Multivariate Students T with

Mean Vector

10.028 0.476 0.548

Covariance Matrix

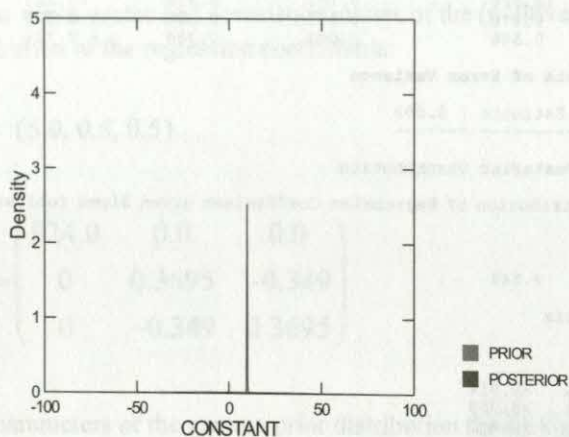
0.010	-0.003	-0.001
-0.003	0.006	-0.005
-0.001	-0.005	0.007

Marginal Distribution of $(1/\text{Sigma})^2$ is

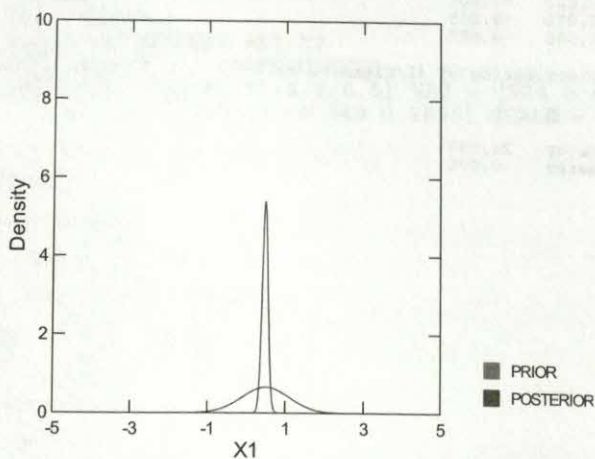
Gamma with

Scale Parameter	24.000
Shape Parameter	0.076

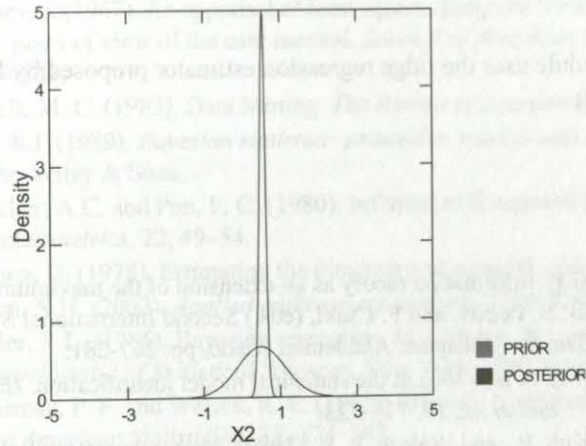
Prior and Posterior Densities of CONSTANT



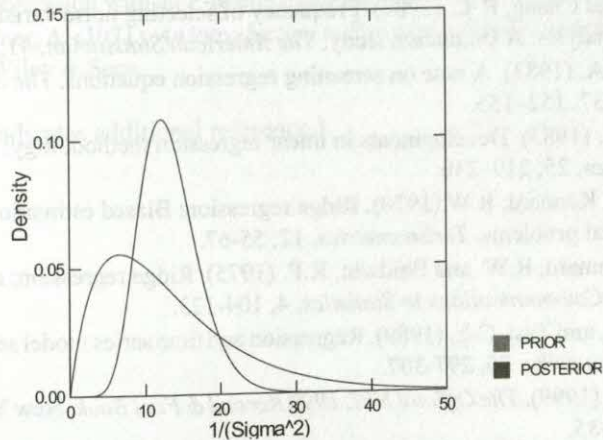
Prior and Posterior Densities of Coefficient of X1



Prior and Posterior Densities of Coefficient of X2



Prior and Posterior Densities of $1/(\text{Sigma}^2)$



Computation

Algorithms

RIDGEREG module uses the ridge regression estimator proposed by Hoerl and Kennard (1970).

References

- *Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. in B. N. Petrov, and F. Csaki, (eds.) Second International Symposium on Information Theory. Budapest: Akademiai Kiado, pp. 267-281.
- *Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC 19, 716-723.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.
- Box, G.E.P., and Tiao, G. C. (1973). Bayesian inference in statistical analysis. Reading, Mass.: Addison-Wesley.
- *Burnham, K.P., and Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.
- *Flack, V. F. and Chang, P. C. (1987). Frequency of selecting noise variables in subset regression analysis: A simulation study. *The American Statistician*, 41, 84-86.
- *Freedman, D. A. (1983). A note on screening regression equations. *The American Statistician*, 37, 152-155.
- *Hocking, R. R. (1983). Developments in linear regression methodology: 1959-82. *Technometrics*, 25, 219-230.
- Hoerl, A.E. and Kennard, R.W.(1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: some simulations, *Communications in Statistics*, 4, 104-123.
- *Hurvich, C.M., and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- *Johnson, R.W. (1999). *The Official NFL 1999 Record & Fact Book*, New York: Workman Publishing, 435.
- Judge, G.G., Griffiths, W.E., Lutkepohl, H., Hill, R.C., and Lee, T. C. (1988). *Introduction to the theory and practice of econometrics*, 2nd ed. New York: John Wiley & Sons, pp. 275-318, pp. 453-454.

- Kvålseth, T. O. (1985). Cautionary note about R^2 . *The American Statistician*, 39, 279.
- Lawless, J.F. and Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics*, A5, 307-323.
- Longley, J. (1967). An appraisal of least squares program for the electronic computer from the point of view of the user manual. *Journal of American Statistical Association*, 62, 819-841.
- *Lovell, M. C. (1983). Data Mining. *The Review of Economics and Statistics*, 65, 1-12.
- Press, S.J. (1989). *Bayesian statistics: principles, models and applications*. New York: John Wiley & Sons.
- *Rencher, A.C. and Pun, F. C. (1980). Inflation of R-squared in best subset regression. *Technometrics*, 22, 49-54.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- *Timm, N.H. (2002). *Applied multivariate analysis*. New York: Springer-Verlag.
- *Trader, R.L. (1986). Bayesian regression. In Johnson, N.L. and Kotz, S. (eds.) *Encyclopedia of Statistical Sciences*, New York: John Wiley & Sons, 7, 677-683.
- *Velleman, P. F. and Welsch, R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician*, 35, 234-242.
- *Weisberg, S. (2005). *Applied linear regression*. 3rd ed. Hoboken, N.J.: Wiley-Interscience.
- *Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, 86, 168-174.
- *Wilkinson, L. and Dallal, G. E. (1982). Tests of significance in forward selection regression with an F-to-enter stopping rule. *Technometrics*, 24, 25-28.
- Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York: John Wiley & Sons.

(* indicates additional reference.)

Linear Models II: Analysis of Variance

Leland Wilkinson and Mark Coward (revised by Sayyad Nisar Badashah and Amol Patil)

SYSTAT handles a wide variety of balanced and unbalanced analysis of variance designs (Speed et al., 1978). The Analysis of Variance (ANOVA) procedure includes all interactions in the model and tests them automatically. Analysis of covariance and the repeated measures designs are a part of the ANOVA feature. Once you have estimated your ANOVA model, it is easy to test the post hoc pairwise differences in means or to test any contrast across cell means, including simple effects.

SYSTAT offers three tests for checking normality: Kolmogorov-Smirnov (Lilliefors), Anderson-Darling, and Shapiro-Wilk test; and Levene's test for homogeneity of variances. You can select any of the three types of sum of squares, Type I, Type II, and Type III, for the analysis.

The ANOVA module provides fifteen tests for pairwise comparisons based on the structure of data and the error rate to be controlled. The pairwise comparison tests are commonly named as post hoc tests; here tests are determined based on the assumptions on variance, viz., equal or unequal variances. One can use post hoc tests after fitting the ANOVA model to check the differences between pairs of means.

The General Linear Model (GLM) procedure is used for randomized block designs (Kutner et al., 2004), incomplete block designs, fractional factorials, Latin square designs (Cochran and Cox, 1957; John, 1971), and analysis of covariance with one or more covariates. GLM also includes repeated measures, split plot, and crossover designs. It includes both univariate and multivariate approaches to repeated measures designs (Bartlett, 1947; Morrison, 2004).

For both ANOVA and GLM, group sizes can be unequal for the combinations of grouping factors; but for repeated measures designs, each subject must have complete data. You can use numeric or character values to code the grouping variables.

You can store results of the analysis (predicted values and residuals) for further study and graphical display. In ANCOVA, you can save adjusted cell means. AIC, AIC (Corrected) and Schwarz's (1978) BIC values are also provided for each fitted model (Burnham and Anderson, 2003). For more information on AIC and Schwarz's BIC in SYSTAT refer to the section "Variable Selection" on page 15 in the chapter on Linear Models in *Statistics II*.

Resampling procedures are available in this feature.

Analysis of Variance in SYSTAT

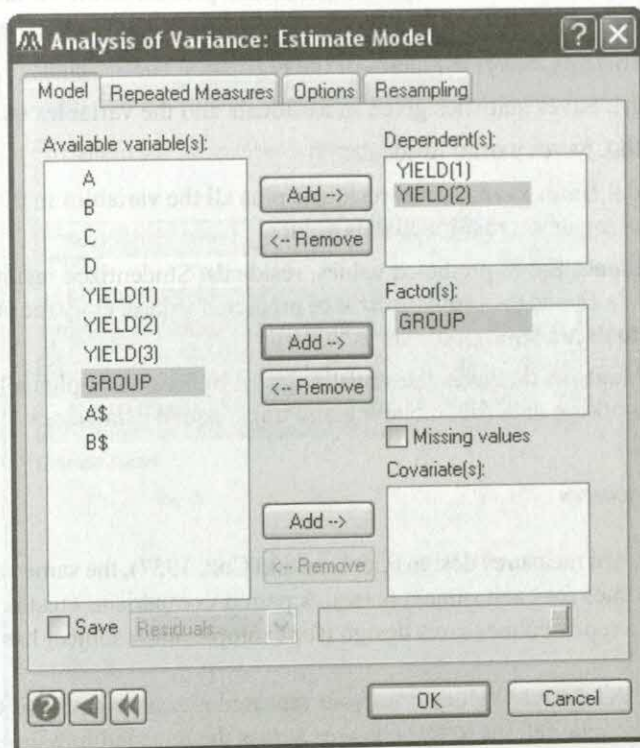
Analysis of Variance: Estimate Model Dialog Box

To obtain an analysis of variance, from the menus choose:

Analyze

Analysis of Variance (ANOVA)

Estimate Model...



Dependent(s). The variable(s) you want to examine. The dependent variable(s) should be continuous and numeric (for example, *INCOME*).

Factor(s). One or more categorical variables (grouping variables) that split your cases into two or more groups.

- **Missing value.** Includes a separate category for cases with a missing value for the variable(s) identified with Factor.

Covariate(s). A covariate is a quantitative independent variable that adds unwanted variability to the dependent variable. An analysis of covariance (ANCOVA) adjusts or removes the variability in the dependent variable due to the covariate (for example, variability in cholesterol level might be removed by using *AGE* as a covariate).

Save. You can save residuals and other data to a new data file. The following alternatives are available:

- **Adjusted.** Saves adjusted cell means from analysis of covariance.

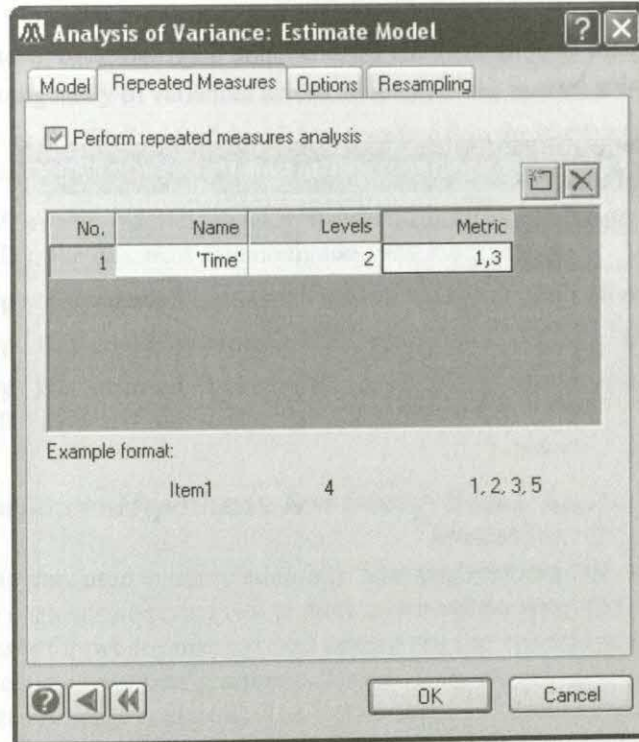
- **Adjusted/Data.** Saves adjusted cell means plus all of the variables in the working data file, including any transformed data values.
- **Coefficients.** Saves estimates of the regression coefficients.
- **Model.** Saves statistics given in Residuals and the variables used in the model.
- **Partial.** Saves partial residuals.
- **Partial/Data.** Saves partial residuals plus all the variables in the working data file, including any transformed data values.
- **Residuals.** Saves predicted values, residuals, Studentized residuals, leverages, Cook's D , and the standard error of predicted values. Only the predicted values and residuals are appropriate for ANOVA.
- **Residuals/Data.** Saves the statistics given by Residuals plus all of the variables in the working data file, including any transformed data values.

Repeated Measures

In a repeated measures design (Cochran and Cox, 1957), the same variable is measured several times for each subject (case). A paired-comparison t test is the most simple form of a repeated measures design (for example, each subject has a before and after measure).

SYSTAT derives values from your repeated measures and uses them in analysis of variance computations to test changes across the repeated measures (within subjects) as well as differences between groups of subjects (between subjects). Tests of the within-subjects values are called polynomial test of order 1, 2, ..., up to k , where k is one less than the number of repeated measures. The first polynomial is used to test linear changes; perform the repeated responses increase (or decrease) around a line with a significant slope. The second polynomial tests whether the responses fall along a quadratic curve, and so on.

To perform repeated measures analysis, click the Repeated Measures tab in Analysis of Variance: Estimate Model dialog box.

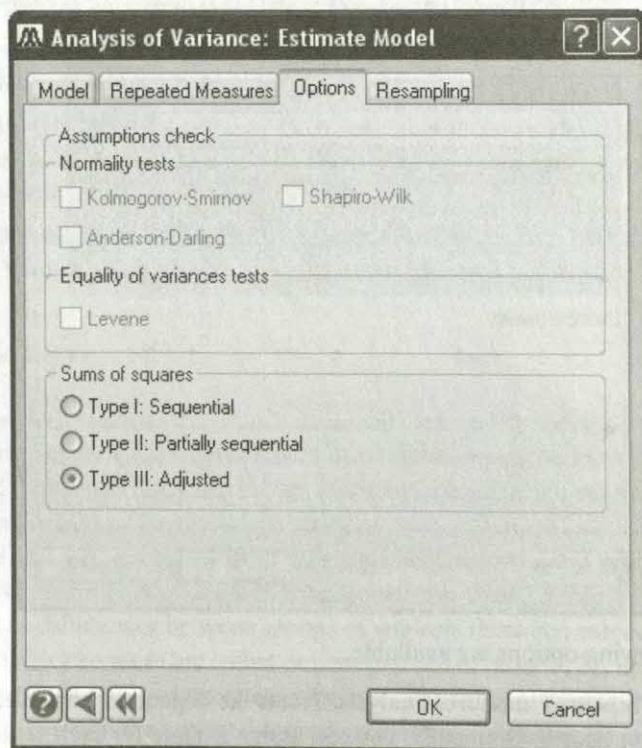


The following options are available:

- **Perform repeated measures analysis.** Treats the dependent variables as a set of repeated measures. Optionally, you can assign a name for each set of repeated measures, specify the number of levels, and specify the metric for unevenly spaced repeated measures.
- **Name.** Name that identifies each set of repeated measures.
- **Levels.** Number of repeated measures in the set. For example, suppose you have three dependent variables that represent measurements at different times, the number of levels is 3.
- **Metric.** Metric that indicates the spacing between unevenly spaced measurements. For example, if measurements were taken at the third, fifth, and ninth weeks, the metric would be 3, 5, 9.

Options

To specify the options, click the Options tab in the Analysis of Variance: Estimate Model dialog box.



Assumptions check. This provides options to check the basic assumptions of ANOVA.

Normality tests. You can use the following normality tests to check the basic statistical assumption of ANOVA, normality of residuals:

- **Kolmogorov-Smirnov.** It is a nonparametric test used for large samples. It is applied to continuous distributions and gives greater importance to the observations in the center than those at the tails.
- **Shapiro-Wilk.** The test provides the Shapiro-Wilk test statistic and p-value for residuals: the smaller the p-value, the worse is the fit.

- **Anderson-Darling.** The Anderson-Darling test is a standard goodness of fit test. It gives greater importance to the observations in the tails than those at the center.

Equality of variances tests. You can use the following equality of variance test to check the homogeneity of variances across all levels of the factors:

- **Levene's.** The Levene's test is less sensitive than the Bartlett test to departures from normality.

Sum of squares. For the model, you can choose a particular type of sum of squares. Type III is the one most commonly used and is the default.

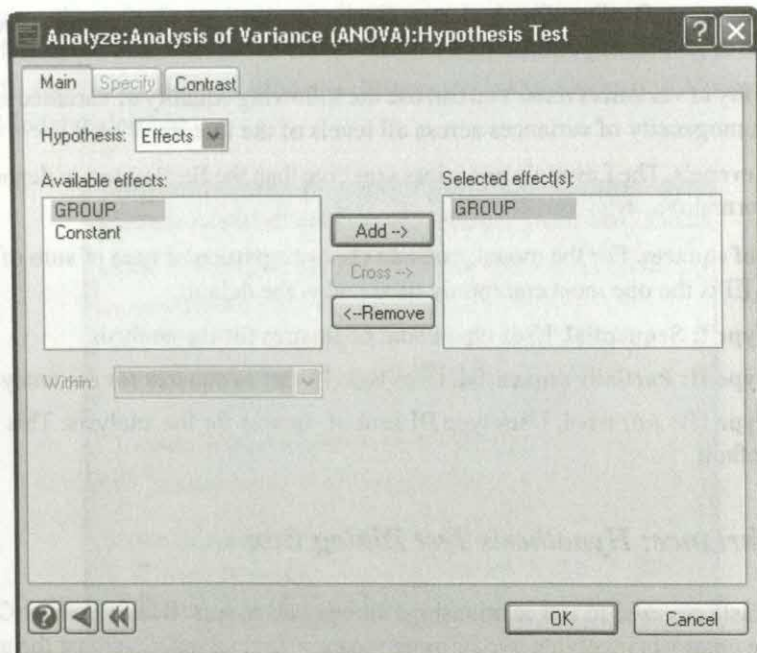
- **Type I: Sequential.** Uses type I sum of squares for the analysis.
- **Type II: Partially sequential.** Uses type II sum of squares for the analysis.
- **Type III: Adjusted.** Uses type III sum of squares for the analysis. This is the default.

Analysis of Variance: Hypothesis Test Dialog Box

Contrasts are used to test relationships among cell means. Use Specify or Contrast to define contrasts involving two or more means—for example, contrast the average responses for two treatment groups against that for a control group; or test if average income increases linearly across cells ordered by education (dropouts, high school graduates, college graduates). The coefficients for the means of the first contrast might be (1, 1, -2) for a contrast of $1 * \text{Treatment A}$ plus $1 * \text{Treatment B}$ minus $2 * \text{Control}$. The coefficients for the second contrast would be (-1, 0, 1). (For more information, see Wilkinson, 1975).

The ANOVA model must be estimated before any hypothesis tests can be performed. To define contrasts among the cell means, from the menus choose:

Analyze
Analysis of Variance (ANOVA)
Hypothesis Test...



Contrasts can be defined across the categories of a grouping factor or across the levels of a repeated measure.

Selected effect(s). Select one or more effects you want to test.

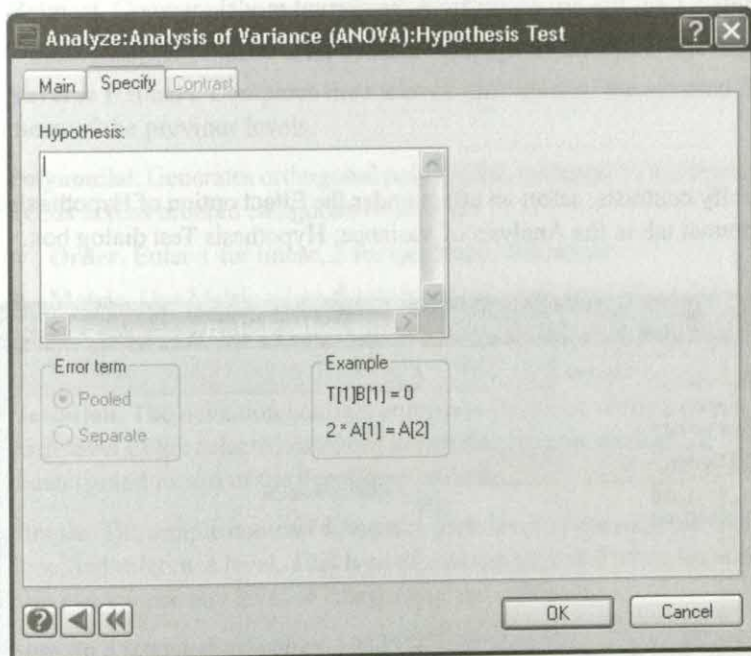
Hypothesis. Select the type of hypothesis. The following choices are available:

- **Model.** Tests for the coefficients of the model. This is the default
- **All.** Select to test all main effects and interactions.
- **Effects.** Select one or more effects you want to test.
- **Specify.** Select Specify to use Specify tab.

Within. Use when specifying a contrast across the levels of repeated measures factor. Select the name assigned to the set of repeated measures in the Repeated Measures tab.

Specify

To specify coefficients for hypothesis tests, select Specify option of Hypothesis in the Analysis of Variance: Hypothesis Test dialog box.



To specify coefficients for a hypothesis test, use cell identifiers. The common hypothesis tests include contrasts across marginal means or tests of simple effects. For a two-way factorial ANOVA design with *DISEASE* (four categories) and *DRUG* (three categories), you could contrast the marginal mean for the first level of drug against the third level by specifying:

$$\text{DRUG}[1] = \text{DRUG}[3]$$

Note that square brackets enclose the value of the category (for example, for *GENDER*\$, specify *GENDER*\$(male)). For the simple contrast of the first and third levels of *DRUG* for the second disease only:

$$\text{DRUG}[1] \text{ DISEASE}[2] = \text{DRUG}[3] \text{ DISEASE}[2]$$

The syntax also allows statements like:

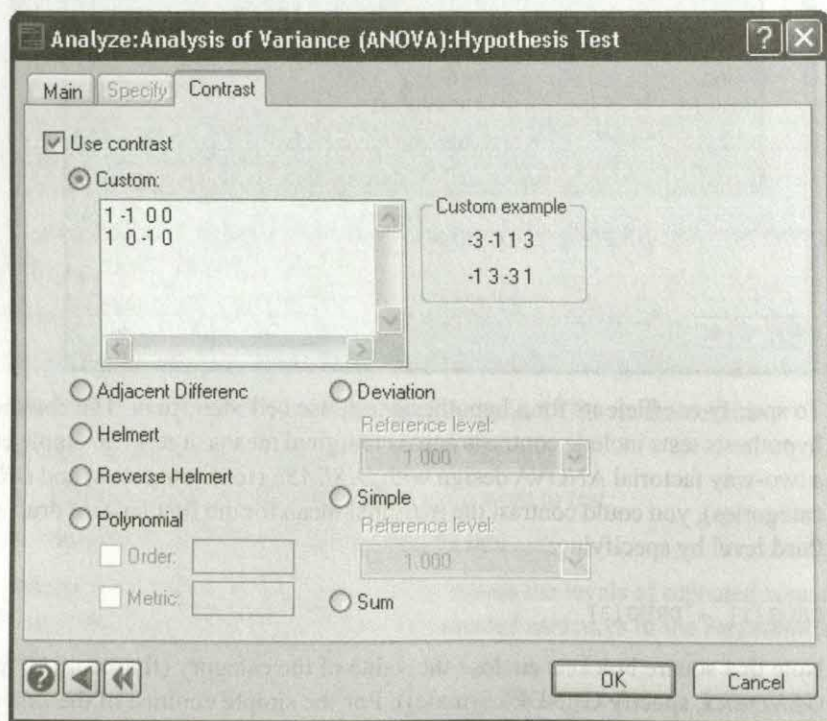
$-3*DRUG[1] - 1*DRUG[2] + 1*DRUG[3] + 3*DRUG[4]$

You have two error term options for hypothesis tests:

- **Pooled.** Uses the error term from the current model.
- **Separate.** Generates a separate variance error term.

Contrast

To specify contrasts, select an effect under the Effect option of Hypothesis and click the Contrast tab in the Analysis of Variance: Hypothesis Test dialog box.



Contrast generates a contrast for a grouping factor or a repeated measures factor. SYSTAT offers eight types of contrasts:

Custom. Enter your own custom coefficients. If your factor has, say, four ordered categories (or levels), you can specify your own coefficients, such as -3 -1 1 3, by typing these values in the Custom text box.

Adjacent difference. Compare each level with its adjacent level of the selected factor.

Helmert. Compares the mean of each level of the selected factor with the mean of the succeeding levels.

Reverse Helmert. Compares the mean of each level of the selected factor with the mean of the previous levels.

Polynomial. Generates orthogonal polynomial contrasts (to test linear, quadratic, cubic trends across ordered categories or levels).

■ **Order.** Enter 1 for linear, 2 for quadratic, and so on.

■ **Metric.** Use Metric when the ordered categories are not evenly spaced. For example, when repeated measures are collected at weeks 2, 4, 8, enter 2, 4, 8 as the metric.

Deviation. The deviation contrast compares the mean of the dependent variable for each level of the selected categorical variable (except a reference level) to the overall mean (grand mean) of the dependent variable.

Simple. The simple contrast compares each level of the selected factor against the specified reference level. This type of contrast is useful when there is a control group. You can choose any level or category as the reference.

Sum. In a repeated measures ANOVA, total the values for each subject.

Analysis of Variance: Pairwise Comparisons Dialog Box

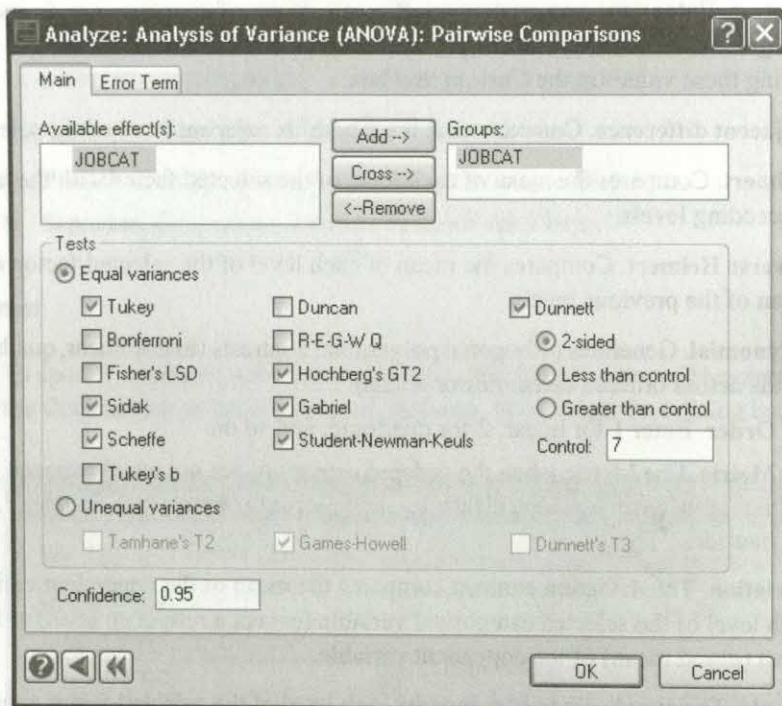
After fitting the model, one can find the treatment pairs which are significantly different, or form several homogeneous sets of treatments with their respective *p-value* by using several multiple comparison tests (mct) offered by SYSTAT under equal or unequal variance assumptions.

To open Pairwise Comparisons dialog box, from menus choose:

Analyze

Analysis of Variance (ANOVA)

Pairwise Comparisons...



Groups. Select the variable that defines the groups.

Tests. There are several post hoc tests to compare the means of the dependent variable for the selected grouping variable.

Equal variances. Tests in this group assume equality of variances across all levels of the grouping variable.

- **Tukey.** Uses the Studentized range distribution to make all pairwise comparisons. This is the default.
- **Bonferroni.** Uses Student's t statistic. It sets the family-wise error rate as $(1 - \text{Confidence})/(\text{Total number of comparisons})$.
- **Fisher's LSD.** Equivalent to multiple t -tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.
- **Sidak.** Uses Student's t statistic for pairwise multiple comparisons.
- **Scheffé.** The significance level of Scheffé's test is designed to allow all possible linear combinations of group means to be tested, not just pairwise comparisons

available in this feature. The result is that Scheffé's test is more conservative than other tests.

- **Tukey's b.** Uses the Studentized range distribution. The critical value is the average of the corresponding values for the Turkey's HSD test and the Student-Newman-Keuls (S-N-K) test.
- **Duncan.** Uses Studentized range distribution. It yields homogeneous subsets of group levels.
- **R-E-G-W Q.** Ryan-Einot-Gabriel-Welsch Q test is a modification of the S-N-K test where the critical values decrease as the range in the set being considered decreases.
- **Hochberg's GT2.** Uses the Studentized maximum modulus distribution
- **Gabriel.** Uses the Studentized maximum modulus distribution. It is equivalent to the GT2 test for balanced ANOVA.
- **Student-Newman-Keuls.** Uses the Studentized range distribution. It yields homogenous subsets of group levels.
- **Dunnett.** The Dunnett test is available only with one-way designs. Dunnett compares a set of treatments against a single control mean that you specify. You can choose from the following three alternative hypotheses: (a) 2-sided (not equal), (b) less than, or (c) greater than the control level. 2-sided is the default.

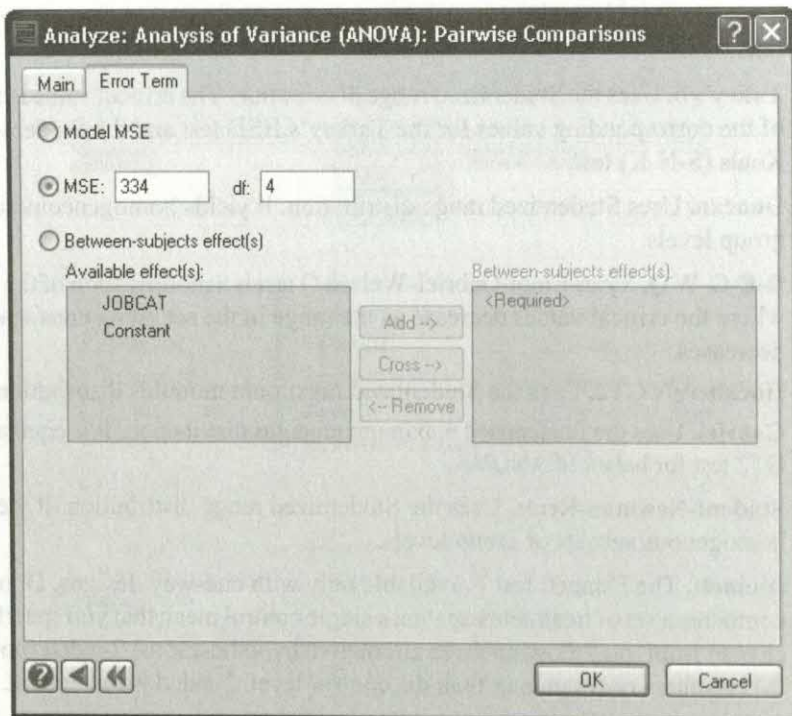
Unequal variances. The following tests do not require the homogeneity of variance assumption. These tests use the Welsch procedure for determining the denominator degrees of freedom.

- **Tamhane's T2.** Uses the Student's *t* distribution. Uses the Sidak inequality to find the alpha level.
- **Games - Howell.** Uses the Studentized range distribution.
- **Dunnett's T3.** Uses the Studentized maximum modulus distribution.

Confidence. Specify confidence level for pairwise comparisons tests. The default value is 0.95.

Error term

To specify the error term, click the Error Term tab in the ANOVA: Pairwise Comparisons dialog box.



You can choose one of the following:

- **Model MSE.** Uses the mean square error (MSE) from the general linear model that you ran.
- **MSE and df.** Uses the mean square error term and degrees of freedom that you specify. Use this option if you know them from a previous model.
- **Between-subjects effect(s).** Select this option to use the main effect error term or the interaction error term in all the tests.

Note: Toggling between the command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through the dialog box then post estimation analysis can be performed through commands, and vice-versa.

Using Commands

```
ANOVA
  USE filename
  CATEGORY varlist/ MISS EFFECT or DUMMY
  DEPEND varlist / REPEAT NAMES
  COVAR varlist
  PLENGTH NONE or SHORT or MEDIUM or LONG
  SAVE filename / ADJUST, COEFFICIENT, MODEL, PARTIAL
                  RESID, DATA 'comment'
  WORK filename / ADJUST, COEFFICIENT, MODEL, PARTIAL
                  RESID, DATA 'comment'
  ESTIMATE / NTEST = KS, SW, AD HTEST = LEVENE
            SS = TYPE1 or TYPE2 or TYPE3 QUICK or NOQUICK
            SAMPLE = BOOT(m,n) or SIMPLE(m,n) or JACK
```

To use ANOVA for analysis of covariance, insert COVARIATE before ESTIMATE.

After estimating a model, use HYPOTHESIS to test its parameters. Begin each test with HYPOTHESIS and end with TEST.

```
HYPOTHESIS
  ALL
  EFFECT varlist or var1*var2 or var1&var2
  WITHIN 'name'
  ERROR value (df) or var or var1*var2 or var1&var2 or matrix
  POST grpvar / TUKEY or BONF=n or LSD or SIDAK or SCHEFFE or
                BTUKEY or DUNCAN or QREGW or GT2 or GABR or
                SNK or GH or T2 or T3 or SEPARATE or
                POOLED or DUNNETT = LT or GT or TWO
                CONTROL = 'levelname'
  CONTRAST / ADJDIFF or POLYNOMIAL, ORDER=n METRIC=m, n,...
            or SUM or DEV[c] or SIMPLE[c] or HEL or RHEL
  SPECIFY / POOLED or SEPARATE
  AMATRIX [matrix]
  CMATRIX [matrix]
  DMATRIX [matrix]
  PAIRWISE / BONFERRONI or SIDAK
  TEST / CONF1 = n
```

Usage Considerations

Types of data. ANOVA requires a rectangular data file.

Print options. If PLENGTH SHORT, the output includes an ANOVA table. The MEDIUM length adds the least-squares means to the output. LONG adds the estimates of the coefficients.

Quick Graphs. ANOVA plots the group means against the groups. .

Saving files. ANOVA can save predicted values, residuals, Studentized residuals, leverages, Cook's *D*, standard error of predicted values, adjusted cell means, and estimates of the coefficients.

BY groups. ANOVA performs separate analyses for each level of any BY variables. However, for Hypothesis Testing, BY groups does not work. You have to resort to Data--> Select Cases commands.

Case frequencies. You can use a FREQUENCY variable to duplicate cases.

Case weights. ANOVA uses a WEIGHT variable, if present, to duplicate cases.

Examples

Example 1

One-Way ANOVA

How does equipment influence typing performance? This example uses a one-way design to compare average typing speed for three groups of typists. Fourteen beginning typists were randomly assigned to three types of machines and given speed tests. The following are their typing speeds in words per minute:

Electric	Plain old	Word processor
52	52	67
47	43	73
51	47	70
49	44	75
53		64

The data are stored in the SYSTAT data file named *TYPING*. The average speeds for the typists in the three groups are 50.4, 46.5, and 69.8 words per minute, respectively. To test the hypothesis that the three samples have the same population average speed, the input is:

```
ANOVA
USE TYPING
CATEGORY EQUIPMNT$
DEPEND SPEED
ESTIMATE
```

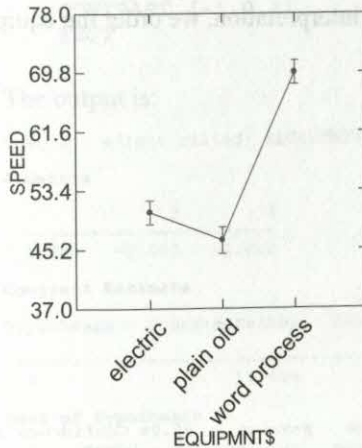
The output is:

Dependent Variable : SPEED
 N : 14
 Multiple R : 0.952
 Squared Multiple R : 0.907

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
EQUIPMNT\$	1469.357	2	734.679	53.520	0.000
Error	151.000	11	13.727		

Least Squares Means



For the dependent variable *SPEED*, SYSTAT reads 14 cases. The multiple correlation (*Multiple R*) for *SPEED* with the two design variables for *EQUIPMNT\$* is 0.952. The square of this correlation (*Squared multiple R*) is 0.907. The grouping structure explains 90.7% of the variability of *SPEED*.

The layout of the ANOVA table is standard in elementary texts; you will find formulas and definitions there. *F-ratio* is the *Mean-Square* for *EQUIPMNT\$* divided by the *Mean-Square* for *Error*. The distribution of the *F-ratio* is sensitive to the assumption of equal population group variances. The *p-value* is the probability of exceeding the *F-ratio* when the group means are equal. The *p-value* printed here is 0.000, so it is less than 0.0005. If the population means are equal, it would be very unusual to find sample means that differ as much as these—you could expect such a large *F-ratio* fewer than five times out of 10,000.

The Quick Graph illustrates this finding. Although the typists using electric and plain old typewriters have similar average speeds (50.4 and 46.5, respectively), the word processor group has a much higher average speed.

Pairwise Mean Comparisons

An analysis of variance indicates whether (at least) one of the groups differs from the others. However, you cannot determine which group(s) differs based on ANOVA results. To examine specific group differences, use post hoc tests.

In this example, we use the Bonferroni method for the typing speed data used in the one-way ANOVA example. As an aid in interpretation, we order the equipment categories from least to most advanced.

The input is:

```
HYPOTHESIS
POST EQUIPMNT$/ BONF
TEST
```

The output is:

```
Post Hoc Test of SPEED
Using least squares means.
Using model MSE of 13.727 with 11 df.
```

Bonferroni Test

EQUIPMNT\$(i)	EQUIPMNT\$(j)	Difference	p-value	95.0% Confidence Interval	
				Lower	Upper
electric	plain old	3.900	0.435	-3.109	10.909
electric	word process	-19.400	0.000	-26.008	-12.792
plain old	word process	-23.300	0.000	-30.309	-16.291

In the first and second rows, you can read differences in average typing speed and corresponding 95% confidence intervals for the group using plain old typewriters. In the third column, row one, you see that the average is 3.9 words per minute fewer than those using electric typewriters; but in the second row, you see that the average is 23.3 minutes fewer than the group using word processors. To see whether these differences are significant, look at the probabilities in the fourth column.

The probability associated with 3.9 is 0.435, so you are unable to detect a difference in performance between the electric and plain old groups. The probabilities in the second and third row are both 0.00, indicating that the word processor group averages significantly more words per minute than the electric and plain old groups.

Contrasts

To compare the differences in the means of levels of a single factor, you can use SYSTAT's CONTRAST command. In this example, suppose you want to contrast 'electric' equipment against 'word processor' equipment, you can use the following commands.

The input is:

```
PLENGTH LONG
HYPOTHESIS
EFFECT EQUIPMNT$
CONTRAST [-1 0 1]
TEST
```

The output is:

Test for effect called: EQUIPMNT\$

A Matrix

	1	2	3
	0.000	-2.000	-1.000

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	19.400	2.343	19.341	19.459

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	940.900	1	940.900	68.542	0.000
Error	151.000	11	13.727		

The model for the above analysis is:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i=1,2,3$ and $j=1(1) n_i$

The parameters μ , α_1 , α_2 and α_3 satisfy the following condition:

$$\sum_{i=1}^3 \alpha_i = 0$$

The contrast in this example is coded as [-1 0 1]. It imposes the restriction on the levels of equipment, which is

$$\alpha_3 - \alpha_1 = 0$$

Using the above assumption, we get

$$(-\alpha_1 - \alpha_2) - \alpha_1 = 0$$

which, in turn, reduces to

$$-2\alpha_1 - \alpha_2 = 0$$

$$\text{i.e. } 0\mu - 2\alpha_1 - \alpha_2 = 0$$

Now, look at the term **A** matrix in the output created by the first and third levels of the equipment. The **A** matrix for the above model is [0 -2 -1]. Notice that the value 0 corresponds to the constant term and -2 and -1 for the first and second design variables in the model. The *Contrast estimate* of the **A** matrix is 19.4, the corresponding *S.E.* is 5.4909 and *95% confidence intervals* are 7.3145, 31.4854.

The *F-ratio* for testing the contrast is 68.542 (*p-value* < 0.0005). Thus you can conclude that there is a significant difference between the first and third levels of equipment.

Similarly the contrast $\alpha_1 - 2\alpha_2 + \alpha_3 = 0$ can be tested by defining the **A** matrix as [0 0 -3].

Example 2

ANOVA Assumptions and Contrasts

An important assumption in analysis of variance is that the population variances are equal—that is, the groups have approximately the same spread. When variances differ markedly, a transformation may remedy the problem. For example, sometimes it helps to take the square root of each value of the outcome variable (or log transform each value) and use the transformed value in the analysis.

In this example, we use a subset of the cases from the *SURVEY2* data file to address the question, “For males, does average income vary with education?” We focus on the following who:

- Did not graduate from high school (*HS dropout*)
- Graduated from high school (*HS grad*)
- Attended some college (*Some college*)

- Graduated from college (*College grad*)
- Have an M.A. or Ph.D. (*Degree +*)

For each male subject (case) in the *SURVEY2* data file, use the variables *INCOME* and *EDUC\$*. The means, standard deviations, and sample sizes for the five groups are shown below:

	HS dropout	HS grad	Some college	College grad	Degree +
mean	\$13,389	\$21,231	\$29,294	\$30,937	\$38,214
sd	10,639	13,176	16,465	16,894	18,230
n	18	39	17	16	14

Visually, as you move across the groups, you see that average income increases. But considering the variability within each group, you might wonder if the differences are significant. Also, there is a relationship between the means and standard deviations—as the means increase, so do the standard deviations. They should be independent. Suppose you take the square root of each income value, there is less variability among the standard deviations, and the relation between the means and standard deviations is weaker:

	HS dropout	HS grad	Some college	College grad	Degree +
mean	3.371	4.423	5.190	5.305	6.007
sd	1.465	1.310	1.583	1.725	1.516

A bar chart for the data will show the effect of the transformation.

The input is:

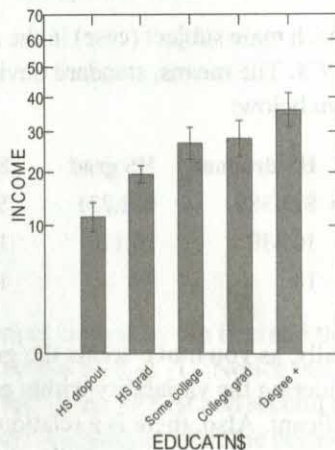
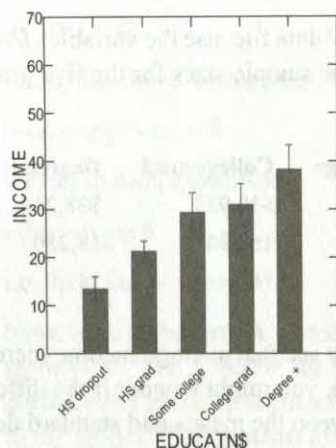
```
USE SURVEY2
SELECT SEX$= 'Male'
RECODE EDUCATN$=EDUCATN / 1,2='HS dropout', 3='HS grad',
                          4='Some college', 5='College,
                          grad' 6,7='Degree +'

CATEGORY EDUCATN$
ORDER EDUCATN$ / SORT = 'HS dropout' 'HS grad' 'Some college',
                        'College grad' 'Degree +'

BEGIN
BAR INCOME * EDUCATN$ / SERROR FILL=.5 LOC=-3IN,0IN
BAR INCOME * EDUCATN$ / SERROR FILL=.35 YPOW=.5,
                        LOC=3IN,0IN

END
```

The output is:



In the chart on the left, you can see a relation between the height of the bars (means) and the length of the error bars (standard errors). The smaller means have shorter error bars than the larger means. After transformation, there is less difference in length among the error bars. The transformation aids in eliminating the dependency between the group and the standard deviation.

To test for differences among the means:

```
ANOVA
LET SQRT_INC = SQR(INCOME)
DEPEND SQRT_INC
CATEGORY EDUCATN$
ESTIMATE/NTEST = KS, SW, AD HTEST = LEVENE
```

The output is:

```
Dependent Variable : SQRT_INC
N : 104
Multiple R : 0.490801
Squared Multiple R : 0.240886
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
EDUCATN\$	68.623582	4	17.155895	7.853793	0.000015
Error	216.256486	99	2.184409		

Test for Homogeneity

	Test Statistic	p-value
Levene's Test	1.005350	0.408535

Test for Normality

	Test Statistic	p-value
K-S Test (Lilliefors)	0.079778	0.099782
Shapiro-Wilk Test	0.989641	0.608196
Anderson-Darling Test	0.424596	>0.15

From the above results of normality tests and the homogeneity test, the assumption of normal residuals is satisfied and the transformed *INCOME* dependent variable fulfills the equal population variance assumption.

The ANOVA table using the transformed income as the dependent variable suggests a significant difference among the four means ($p\text{-value} < 0.0005$).

Tukey Pairwise Mean Comparisons

Which means differ? This example uses the Tukey method to identify significant differences in pairs of means. Hopefully, you reach the same conclusions using either the Tukey or Bonferroni methods. However, when the number of comparisons is very large, the Tukey procedure may be more sensitive in detecting differences; when the number of comparisons is small, Bonferroni may be more sensitive.

The input is:

```
HYPOTHESIS
POST EDUCATN$ / TUKEY
TEST
```

The output is:

```
Post Hoc Test of SQRT_INC
Using least squares means.

Using model MSE of 2.184409 with 99 df.
```


Tukey's Honestly-Significant-Difference Test

EDUCATN\$ (i)	EDUCATN\$ (j)	Difference	p-value	95.0% Confidence Interval Lower	Interval Upper
HS dropout	HS grad	-1.051736	0.099508	-2.221989	0.118517
HS dropout	Some college	-1.819060	0.003917	-3.208003	-0.430118
HS dropout	College grad	-1.934562	0.002209	-3.345650	-0.523474
HS dropout	Degree +	-2.635771	0.000028	-4.099247	-1.172295
HS grad	Some college	-0.767324	0.387284	-1.960895	0.426247
HS grad	College grad	-0.882826	0.267967	-2.102096	0.336445
HS grad	Degree +	-1.584035	0.007454	-2.863571	-0.304498
Some college	College grad	-0.115502	0.999430	-1.545987	1.314984
Some college	Degree +	-0.816711	0.544944	-2.298899	0.665478
College grad	Degree +	-0.701209	0.694029	-2.204170	0.801752

The layout of the output panels for the Tukey method is the same as that for the Bonferroni method. Look first at the probabilities in the fourth column. Four of the probabilities indicate significant differences (they are less than 0.05). In the third column, row 2, 3 and 4, the average income for high school dropouts differs from those with some college ($p\text{-value} = 0.003$), from college graduates ($p\text{-value} = 0.002$), and also from those with advanced degrees ($p\text{-value} < 0.0005$). The seventh row shows that the differences between those with advanced degrees and the high school graduates are significant ($p\text{-value} = 0.007$).

Contrasts

In this example, the five groups are ordered by their level of education, so you use these coefficients to test linear and quadratic contrasts:

Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2

Then you ask, "Is there a linear increase in average income across the five ordered levels of education?" "A quadratic change?"

The input is:

HYPOTHESIS

NOTE 'Test of linear contrast',
'across ordered group means'

EFFECT EDUCATN\$

CONTRAST [-2 -1 0 1 2]

TEST

HYPOTHESIS

NOTE 'Test of quadratic contrast',
'across ordered group means'

EFFECT EDUCATN\$

CONTRAST [2 -1 -2 -1 2]

TEST

SELECT

The output is:

Test of linear contrast
across ordered group means

Test for effect called: EDUCATN\$

A Matrix

	1	2	3	4	5
0.000000	-4.000000	-3.000000	-2.000000	-1.000000	

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	6.154368	1.141086	6.125841	6.182895

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	63.542478	1	63.542478	29.089094	0.000000
Error	216.256486	99	2.184409		

Test of quadratic contrast
across ordered group means

Test for effect called: EDUCATN\$

A Matrix

	1	2	3	4	5
0.000000	0.000000	-3.000000	-4.000000	-3.000000	

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	-1.352877	1.347611	-1.386568	-1.319187

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval Lower	Upper
A	-1.352877	1.347611	-1.386568	-1.319187

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	2.201515	1	2.201515	1.007831	0.317870
Error	216.256486	99	2.184409		

The *F*-ratio for testing the linear contrast is 29.089 (p -value < 0.0005); for testing the quadratic contrast, it is 1.008 (p -value = 0.318). Thus, you can report that there is a highly significant linear increase in average income across the five levels of education and that you have not found a quadratic component in this increase.

Example 3

Two-Way ANOVA

Consider the following two-way analysis of variance design from Afifi and Azen (1972), cited in Kutner (1974). The dependent variable, *SYSINCR*, is the change in systolic blood pressure after administering one of four different drugs to patients with one of three different diseases. Patients were assigned randomly to one of the possible drugs. The data are stored in the SYSTAT file *AFIFI*.

To obtain a least-squares two-way analysis of variance, the input is:

```
ANOVA
USE AFIFI
CATEGORY DRUG DISEASE
DEPEND SYSINCR
SAVE MYRESIDS / RESID DATA
ESTIMATE
```

Because this is a factorial design, ANOVA automatically generates an interaction term (*DRUG * DISEASE*).

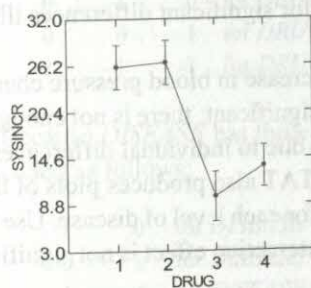
The output is:

Dependent Variable	SYSINCR
N	58
Multiple R	0.675
Squared Multiple R	0.456

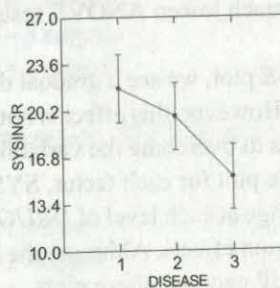
Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DRUG	2997.472	3	999.157	9.046	0.000
DISEASE	415.873	2	207.937	1.883	0.164
DRUG*DISEASE	707.266	6	117.878	1.067	0.396
Error	5080.817	46	110.453		

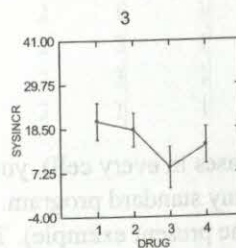
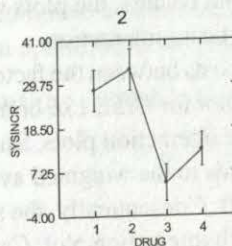
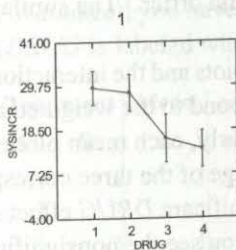
Least Squares Means



Least Squares Means



Least Squares Means



In two-way ANOVA, begin by examining the interaction. If the interaction is significant, you must condition your conclusions about a given factor's effects on the level of the other factor. The *DRUG * DISEASE* interaction is not significant ($p\text{-value} = 0.396$), so shift your focus to the main effects.

The *DRUG* effect is significant ($p\text{-value} < 0.0005$), but the *DISEASE* effect is not ($p\text{-value} = 0.164$). Thus, at least one of the drugs differs from the others with respect to blood pressure change, but blood pressure change does not vary significantly across diseases.

For each factor, SYSTAT produces a plot of the average value of the dependent variable for each level of the factor. For the *DRUG* plot, drugs 1 and 2 yield similar average blood pressure changes. However, the average blood pressure change for drugs 3 and 4 are much lower. ANOVA tests for significant differences illustrated in this plot.

For the *DISEASE* plot, we see a gradual decrease in blood pressure change across the three diseases. However, this effect is not significant; there is not enough variation among these means to overcome the variation due to individual differences.

In addition to the plot for each factor, SYSTAT also produces plots of the average blood pressure change at each level of *DRUG* for each level of disease. Use these plots to illustrate interaction effects. Although the interaction effect is not significant in this example, we can still examine these plots.

In general, we see a decline in blood pressure change across drugs. (Keep in mind that the drugs are only artificially ordered. We could reorder the drugs, and although the ANOVA results would not change, the plots would differ.) The similarity of the plots illustrates the nonsignificant interaction.

A close correspondence exists between the factor plots and the interaction plots. The means plotted in the factor plot for *DISEASE* correspond to the weighted average of the four points in each of the interaction plots. Similarly, each mean plotted in the *DRUG* factor plot corresponds to the weighted average of the three corresponding points across interaction plots. Consequently, the significant *DRUG* effect can be seen in the differing means in each interaction plot. Can you see the nonsignificant *DISEASE* effect in the interaction plots?

Least-Squares ANOVA

If you have an orthogonal design (equal number of cases in every cell), you will find that the ANOVA table is the same one you get with any standard program. SYSTAT can handle non-orthogonal designs, however (as in the present example). To

understand the sources for sum of squares, you must know something about least-squares ANOVA.

As with one-way ANOVA, your specifying factor levels causes SYSTAT to create dummy variables out of the classifying input variable. SYSTAT creates one fewer dummy variable than the categories specified.

Coding of the dummy variables is the classic analysis of variance parameterization, in which the sum of effects estimated for a classifying variable is 0 (Scheffé, 1959). In our example, *DRUG* has four categories; therefore, SYSTAT creates three dummy variables with the following scores for subjects at each level:

1	0	0	for <i>DRUG</i> = 1 subject
0	1	0	for <i>DRUG</i> = 2 subjects
0	0	1	for <i>DRUG</i> = 3 subjects
-1	-1	-1	for <i>DRUG</i> = 4 subjects

Because *DISEASE* has three categories, SYSTAT creates two dummy variables to be coded as follows:

1	0	for <i>DISEASE</i> = 1 subject
0	1	for <i>DISEASE</i> = 2 subjects
-1	-1	for <i>DISEASE</i> = 3 subjects

Now, because there are no continuous predictors in the model (unlike the analysis of covariance), you have a complete design matrix of dummy variables as follows (*DRUG* is labeled with an *a*, *DISEASE* with a *b*, and the grand mean with an *m*):

Treatment		Mean	DRUG			DISEASE		Interaction					
A	B		a1	a2	a3	b1	b2	a1b1	a1b2	a2b1	a2b2	a3b1	a3b2
1	1	1	1	0	0	1	0	1	0	0	0	0	0
1	2	1	1	0	0	0	1	0	1	0	0	0	0
1	3	1	1	0	0	-1	-1	-1	-1	0	0	0	0
2	1	1	0	1	0	1	0	0	0	1	0	0	0
2	2	1	0	1	0	0	1	0	0	0	1	0	0
2	3	1	0	1	0	-1	-1	0	0	-1	-1	0	0
3	1	1	0	0	1	1	0	0	0	0	0	1	0

3	2	1	0	0	1	0	1	0	0	0	0	0	1
3	3	1	0	0	1	-1	-1	0	0	0	0	-1	-1
4	1	1	-1	-1	-1	1	0	-1	0	-1	0	-1	0
4	2	1	-1	-1	-1	0	1	0	-1	0	-1	0	-1
4	3	1	-1	-1	-1	-1	-1	1	1	1	1	1	1

This example is used to explain how SYSTAT gets an error term for the ANOVA table. Because it is a least-squares program, the error term is taken from the residual sum of squares in the regression onto the above dummy variables. For non-orthogonal designs, this choice is identical to that produced by GLM with Type III sum of squares. These, in general, will be the hypotheses you want to test on unbalanced experimental data. You can construct other types of sum of squares by using an **A** matrix or by running your ANOVA model using the Stepwise options in GLM. Consult the references and or Chapter 1: Linear Models of *Statistics II* if you do not already know what these sum of squares mean.

Simple and deviation contrasts

It is evident that only the main effect for *DRUG* is significant; therefore, you might want to test some specified contrasts on the *DRUG* effects. To compare a specified drug level with other drug levels, we can use the SIMPLE contrast and to compare each drug level with the mean of other *DRUG* levels, we can use DEVIATION contrast.

The input is:

```
PLENGTH LONG
HYPOTHESIS
EFFECT DRUG
CONTRAST / DEVIATION[4]
TEST
```

```
HYPOTHESIS
EFFECT DRUG
CONTRAST / SIMPLE[4]
TEST
```


The following are the results of the above hypothesis tests:

Test for effect called: DRUG

A Matrix

	1	2	3	4	5
1	0.000	-1.333	0.000	0.000	0.000
2	0.000	0.000	-1.333	0.000	0.000
3	0.000	0.000	0.000	-1.333	0.000

A Matrix

	6	7	8	9	10
1	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000

A Matrix

	11	12
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000

Contrast Estimate

Hypothesis	Estimate (AB)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A1	-9.380	3.202	-9.460	-9.300
A2	-10.128	3.202	-10.208	-10.048
A3	12.287	3.474	12.200	12.374

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	948.050	1	948.050	8.583	0.005
A2	1105.320	1	1105.320	10.007	0.003
A3	1381.771	1	1381.771	12.510	0.001
A	2997.472	3	999.157	9.046	0.000
Error	5080.817	46	110.453		

Note that simultaneously and marginally each level of *DRUG* differs significantly from the mean of the other *DRUG* levels.

Test for effect called: DRUG

A Matrix

	1	2	3	4	5
1	0.000	2.000	1.000	1.000	0.000
2	0.000	1.000	2.000	1.000	0.000
3	0.000	1.000	1.000	2.000	0.000

A Matrix

	6	7	8	9	10
1	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000

A Matrix

	11	12
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000

Contrast Estimate

Hypothesis	Estimate(AB)	Standard Error	95.0% Confidence Interval Lower	Upper
A1	12.450	3.811	12.355	12.545
A2	13.011	3.811	12.916	13.106
A3	-3.800	4.070	-3.902	-3.698

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	1178.892	1	1178.892	10.673	0.002
A2	1287.550	1	1287.550	11.657	0.001
A3	96.267	1	96.267	0.872	0.355
A	2997.472	3	999.157	9.046	0.000
Error	5080.817	46	110.453		

Observe that A3 ($p\text{-value} = 0.355397$) is insignificant, that is, only the third and the fourth *DRUG* levels are not significantly different.

Custom Contrasts

A simple way to test *DRUG* contrasts would be to use the Bonferroni method to test all pairwise comparisons (Miller, 1985) of marginal drug means. However, to compare three or more means, you must specify the particular contrasts of interest. Here, we compare the first and third drugs, the first and fourth drugs, and the first two drugs with the last two drugs.

The input is:

```

HYPOTHESIS
EFFECT DRUG
CONTRAST [1 0 -1 0]
TEST
HYPOTHESIS
EFFECT DRUG
CONTRAST [1 0 0 -1]
TEST
HYPOTHESIS
EFFECT DRUG
CONTRAST [1 1 -1 -1]
TEST

```

You need four numbers in each contrast because *DRUG* has four levels. You cannot use CONTRAST to specify coefficients for interaction terms. It creates an A matrix only for main effects. The following are the results of the above hypothesis tests:

Test for effect called: DRUG

A Matrix

1	2	3	4	5
0.000	1.000	0.000	-1.000	0.000

A Matrix

6	7	8	9	10
0.000	0.000	0.000	0.000	0.000

A Matrix

11	12
0.000	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	1697.545	1	1697.545	15.369	0.000
Error	5080.817	46	110.453		

Test for effect called: DRUG

A Matrix

1	2	3	4	5
0.000	2.000	1.000	1.000	0.000

A Matrix

6	7	8	9	10
0.000	0.000	0.000	0.000	0.000

A Matrix

11	12
0.000	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	1178.892	1	1178.892	10.673	0.002
Error	5080.817	46	110.453		

Test for effect called: DRUG

A Matrix

1	2	3	4	5
0.000	2.000	2.000	0.000	0.000

A Matrix

6	7	8	9	10
0.000	0.000	0.000	0.000	0.000

A Matrix

11	12
0.000	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	2982.934	1	2982.934	27.006	0.000
Error	5080.817	46	110.453		

Notice the **A** matrices in the output. SYSTAT automatically takes into account the degree of freedom lost in the design coding. Also, notice that you do not need to normalize contrasts or rows of the **A** matrix to unit vector length, as in some ANOVA programs. If you use (2 0 -2 0) or (0.707 0 -0.707 0) instead of (1 0 -1 0), you get the same sum of squares.

For the comparison of the first and third drugs, the *F-ratio* is 15.369 (*p-value* < 0.0005), indicating that these two drugs differ. Looking at the Quick Graphs produced earlier, we see that the change in blood pressure was much smaller for the third drug.

Notice that in the **A** matrix created by the contrast of the first and fourth drugs, you get (2 1 1) in place of the three design variables corresponding to the appropriate columns of the **A** matrix. Because you selected the reduced form for coding of design variables in which sums of effects are 0, you have the following restriction for the *DRUG* effects:

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$$

where each value is the effect for that level of *DRUG*. This means that:

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

and the contrast *DRUG*(1) - *DRUG*(4) is equivalent to:

$$\alpha_1 - [-(\alpha_1 + \alpha_2 + \alpha_3)] = 0$$

which is:

$$2\alpha_1 + \alpha_2 + \alpha_3 = 0$$

For the final contrast, SYSTAT transforms the (1 1 -1 -1) specification into contrast coefficients of (2 2 0) for the dummy coded variables. The p -value (< 0.0005) indicates that the first two drugs differ from the last two drugs.

Simple Effects

You can do simple contrasts between drugs within levels of disease (although the lack of a significant *DRUG* * *DISEASE* interaction does not justify it). To show how it is done, consider a contrast between the first and third levels of *DRUG* for the first *DISEASE* only. You must specify the contrast in terms of the cell means. Use the terminology:

$$\text{MEAN (DRUG index, DISEASE index)} = M\{i, j\}$$

You want to contrast cell means $M\{1,1\}$ and $M\{3,1\}$. These are composed of:

$$M\{1, 1\} = \mu + \alpha_1 + \beta_1 + \alpha\beta_{11}$$

$$M\{3, 1\} = \mu + \alpha_3 + \beta_1 + \alpha\beta_{31}$$

Therefore the difference between the two means is:

$$M\{1, 1\} - M\{3, 1\} = \alpha_1 - \alpha_3 + \alpha\beta_{11} - \alpha\beta_{31}$$

Now, suppose you consider the coding of the variables, you can construct an **A** matrix that picks up the appropriate columns of the design matrix. Here are the column labels of the design matrix (*a* means *DRUG* and *b* means *DISEASE*) to serve as a column ruler over the **A** matrix specified in the hypothesis.

m	a1	a2	a3	b1	b2	a1b1	a1b2	a2b1	a2b2	a3b1	a3b2
0	1	0	-1	0	0	1	0	0	0	-1	0

The input is:

```

HYPOTHESIS
AMATRIX [0 1 0 -1 0 0 1 0 0 0 -1 0]
TEST

```


The output is:

A Matrix

1	2	3	4	5
0.000	1.000	0.000	-1.000	0.000

A Matrix

6	7	8	9	10
0.000	1.000	0.000	0.000	0.000

A Matrix

11	12
-1.000	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	338.000	1	338.000	3.060	0.087
Error	5080.817	46	110.453		

After you understand how SYSTAT codes design variables and how the model sentence orders them, you can take any standard ANOVA text like Winer, Brown and Michels (1991) or Scheffé (1959) and construct an A matrix for any linear contrast.

Contrasting Marginal and Cell Means

Now look at how to contrast cell means directly without being concerned about how they are coded. Test the first level of *DRUG* against the third (contrasting the marginal means).

The input is:

```
HYPOTHESIS
SPECIFY DRUG[1] = DRUG[3]
TEST
```

To contrast the first against the fourth:

```
HYPOTHESIS
SPECIFY DRUG[1] = DRUG[4]
TEST
```

Finally, here is the simple contrast of the first and third levels of *DRUG* for the first *DISEASE* only:

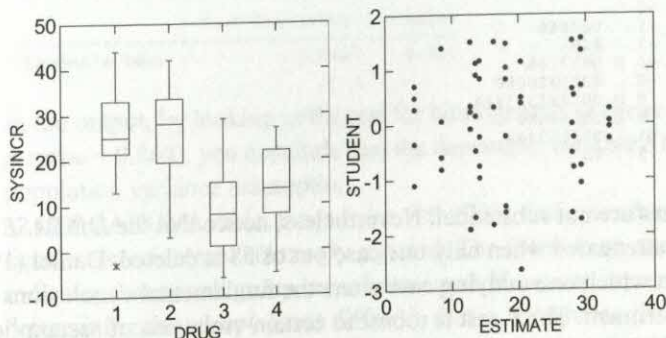
```
HYPOTHESIS
SPECIFY DRUG[1] DISEASE[1] = DRUG[3] DISEASE[1]
TEST
```

Screening Results

Let's examine the *AFIFI* data in more detail. To analyze the residuals to examine the ANOVA assumptions, first plot the residuals against estimated values (cell means) to check for homogeneity of variance. Use the Studentized residuals to reference them against a *t* distribution. In addition, stem-and-leaf plots of the residuals and boxplots of the dependent variable aid in identifying outliers.

The input is:

```
ANOVA
USE AFIFI
CATEGORY DRUG DISEASE
DEPEND SYSINCR
SAVE MYRESIDS / RESID DATA
ESTIMATE
DENSITY SYSINCR * DRUG / BOX
USE MYRESIDS
PLOT STUDENT*ESTIMATE / SYM=1 FILL=1
STEM STUDENT
```



```

Dependent Variable : SYSINCR
N : 58
Multiple R : 0.675
Squared Multiple R : 0.456

```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DRUG	2997.472	3	999.157	9.046	0.000
DISEASE	415.873	2	207.937	1.883	0.164
DRUG*DISEASE	707.266	6	117.878	1.067	0.396
Error	5080.817	46	110.453		

The plots suggest the presence of an outlier. The smallest value in the stem-and-leaf plot seems to be out of line. A t statistic value of -2.647 corresponds to $p\text{-value} < 0.01$, and you would not expect a value this small to show up in a sample of only 58 independent values. In the scatterplot, the point corresponding to this value appears at the bottom and badly skews the data in its cell (which happens to be *DRUG1*, *DISEASE3*). The outlier in the first group also clearly stands out in the boxplot. To see the effect of this outlier, delete the observation with the outlying Studentized residual.

Then, run the analysis again:

Stem and Leaf Plot of Variable: STUDENT, N = 58

```

Minimum : -2.647
Lower Hinge : -0.761
Median : 0.101
Upper Hinge : 0.698
Maximum : 1.552

-2 6
-2
-1 987666
-1 410
-0 H 9877765
-0 4322220000
0 M 001222333444
0 H 55666888
1 011133444
1 55

```

The differences are not substantial. Nevertheless, notice that the *DISEASE* effect is substantially attenuated when only one case out of 58 is deleted. Daniel (1960) gives an example in which one outlying case alters the fundamental conclusions of a designed experiment. The F-test is robust to certain violations of assumptions, but factorial ANOVA is not robust against outliers. You should routinely do these plots for ANOVA.

Example 4

Pairwise comparisons

An analysis of variance indicates whether (at least) one of the groups differs from the others. However, you cannot determine which group(s) differ(s) based on ANOVA results. To examine specific group differences, use post hoc tests.

In this example, we use the *AFIFI* data to test for the difference between *DRUG* levels.

The input is:

```
ANOVA
USE AFIFI
DEPEND SYSINCR
CATEGORY DRUG
ESTIMATE/HTEST = LEVENE
```

The output is:

```
Dependent Variable | SYSINCR
N | 58
Multiple R | 0.579
Squared Multiple R | 0.335
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DRUG	3133.239	3	1044.413	9.086	0.000
Error	6206.917	54	114.943		

Test for Homogeneity

	Test Statistic	p-value
Levene's Test	0.246	0.864

In the output, by looking at the test for homogeneity (*Levene's test* statistics = 0.246, *p-value* = 0.864), you conclude that the dependent variable *SYSINCR* fulfills the equal population variance assumption.

In the ANOVA table, the *p-value* (<0.0005), indicates that the null hypothesis of equal means is overwhelmingly rejected. The F-test in an analysis of variance only indicates that not all group means are equal. However, one may be interested in determining the group(s) that differ(s), that is, the groups that are responsible for the rejection of the null hypothesis of equal means. To examine specific group differences and perhaps to order the groups according to their means, one may use the following post hoc tests.

The input is:

```
HYPOTHESIS
POST DRUG / SNK
TEST/CONFI=0.95
```

The output is:

Post Hoc Test of SYSINCR
Using least squares means.
Using model MSE of 114.943 with 54 df.

Student-Newman-Keuls Test

SubGroup	DRUG	Group Mean	Group Size	p-value
1	3	8.750	15.000	0.241
	4	13.500	15.000	
2	2	25.533	12.000	0.895
	1	26.067	16.000	

* This test controls family-wise error rate under the complete null hypothesis but not under partial null hypothesis.

The Student-Newman-Keuls test displays homogeneous subset numbers, factor levels, ordered group means, group size, and *p-value* for each subset of the treatments under consideration. The above output shows that groups 3 and 4 belong to the same homogeneous subset (the corresponding *p-value* is 0.241), whereas the rest of the groups belong to another subset (*p-value* is 0.895).

Example 5

Unbalanced ANOVA

To test the effect of *DRUG*, *DISEASE*, and *DRUG * DISEASE* interaction on the response variable, three different types of sum of squares are used.

The input is:

```
ANOVA
USE AFIFI
DEPEND SYSINCR
CATEGORY DRUG DISEASE
ESTIMATE/SS = TYPE1
```

```
ANOVA
  DEPEND SYSINCR
  CATEGORY DRUG DISEASE
  ESTIMATE/SS = TYPE2
```

```
ANOVA
  DEPEND SYSINCR
  CATEGORY DRUG DISEASE
  ESTIMATE/SS = TYPE3
```

The output is:

```
Dependent Variable | SYSINCR
N                  |      58
Multiple R         |    0.675
Squared Multiple R |    0.456
```

Analysis of Variance

Source	Type I SS	df	Mean Squares	F-ratio	p-value
DRUG	3133.239	3	1044.413	9.456	0.000
DISEASE	418.834	2	209.417	1.896	0.162
DRUG*DISEASE	707.266	6	117.878	1.067	0.396
Error	5080.817	46	110.453		

Analysis of Variance

Source	Type II SS	df	Mean Squares	F-ratio	p-value
DRUG	3063.433	3	1021.144	9.245	0.000
DISEASE	418.834	2	209.417	1.896	0.162
DRUG*DISEASE	707.266	6	117.878	1.067	0.396
Error	5080.817	46	110.453		

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DRUG	2997.472	3	999.157	9.046	0.000
DISEASE	415.873	2	207.937	1.883	0.164
DRUG*DISEASE	707.266	6	117.878	1.067	0.396
Error	5080.817	46	110.453		

Note the differences between the three types of sum of squares. The Type I sum of squares for *DRUG* essentially tests the differences between the expected values of the arithmetic mean response for different drugs; testing the effect of the disease is not taken into account. The Type II sum of squares for *DRUG* measures the difference between the arithmetic means for each drug after adjusting for the disease. The Type III sum of squares measures the difference between the least-squares means for drug levels.

No matter which sum of squares you use, the above analysis shows significant differences among the four drugs, while the *DISEASE* effect and the *DRUG*DISEASE* interaction are not significant.

Example 6

Single-Degree-of-Freedom Designs

The data in the *REACT* file involve yields of a chemical reaction under various combinations of four binary factors (*A*, *B*, *C*, and *D*). Two reactions were observed under each combination of experimental factors, so the number of cases per cell is two. To analyze the data in a four-way ANOVA, the input is:

```
ANOVA
USE REACT
CATEGORY A, B, C, D
DEPEND YIELD
ESTIMATE
```

You can see the advantage of ANOVA over GLM when you have several factors; you have to select only the main effects. With GLM, you have to specify the interactions and identify which variables are categorical (that is, *A*, *B*, *C*, and *D*). The following example is the full model using GLM:

```
MODEL YIELD = CONSTANT + A + B + C + D +,
              A*B + A*C + A*D + B*C + B*D + C*D +,
              A*B*C + A*B*D + A*C*D + B*C*D +,
              A*B*C*D
```

The output is:

```
Dependent Variable | YIELD
N                  | 32
Multiple R         | 0.755
Squared Multiple R | 0.570
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
A	369800.000	1	369800.000	4.651	0.047
B	1458.000	1	1458.000	0.018	0.894
C	5565.125	1	5565.125	0.070	0.795
D	172578.125	1	172578.125	2.170	0.160
A*B	87153.125	1	87153.125	1.096	0.311
A*C	137288.000	1	137288.000	1.727	0.207
A*D	328860.500	1	328860.500	4.136	0.059
B*C	61952.000	1	61952.000	0.779	0.390
B*D	3200.000	1	3200.000	0.040	0.844
C*D	3160.125	1	3160.125	0.040	0.844
A*B*C	81810.125	1	81810.125	1.029	0.326
A*B*D	4753.125	1	4753.125	0.060	0.810
A*C*D	415872.000	1	415872.000	5.230	0.036
B*C*D	4.500	1	4.500	0.000	0.994
A*B*C*D	15051.125	1	15051.125	0.189	0.669
Error	1272247.000	16	79515.438		

The output shows a significant main effect for the first factor (*A*) plus one significant interaction (*A*C*D*).

Assessing Normality

Let's look at the study more closely. Because this is a single degree of freedom study (a 2^n factorial), each effect estimate is normally distributed if the usual assumptions for the experiment are valid. All of the effects estimates, except the constant, have zero mean and common variance (because dummy variables were used in their computation). Thus, you can compare them to a normal distribution. SYSTAT remembers your last selections.

The input is:

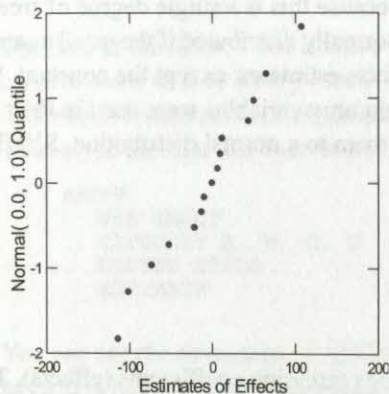
```
SAVE EFFECTS / COEF
ESTIMATE
```

This reestimates the model and saves the regression coefficients (effects). The file has one case with 16 variables (*CONSTANT* plus 15 effects). The effects are labeled $X(1)$, $X(2)$, and so on because they are related to the dummy variables, not the original variables A , B , C , and D . Let's transpose this file into a new file containing only the 15 effects and create a probability plot of the effects.

The input is:

```
USE EFFECTS
DROP CONSTANT
TRANPOSE
PLOT col(1) / FILL=1 SYMBOL=1,
XLABEL="Estimates of Effects"
```


The output is:



These effects are indistinguishable from a random normal variable. They plot almost on a straight line. What does it mean for the study and for the significant F-test?

It is time to reveal that the data were produced by a random number generator.

- If you are doing a factorial analysis of variance, the *p-value* you see on the output are not adjusted for the number of factors. If you do a three-way design, look at seven tests (excluding the constant). For a four-way design, examine 15 tests. Out of 15 F-test on random data, expect to find at least one test approaching significance. You have two significant and one almost significant, which is not far out of line. The probabilities for each separate F-test need to be corrected for the experimentwise error rate. Some authors devote entire chapters to fine distinctions between multiple comparison procedures and then illustrate them within a multifactorial design not corrected for the experimentwise error rate just demonstrated. Remember that a factorial design is a multiple comparison. If you have a single-degree-of-freedom study, use the procedure you used to draw a probability plot of the effects. Any effect that is really significant will become obvious.
- If you have a factorial study with more degrees of freedom on some factors, use the Bonferroni critical value for deciding which effects are significant. It guarantees that the Type I error rate for the study will be no greater than the level you choose. In the above example, this value is $0.05 / 15$ (that is, 0.003).
- Multiple F-tests based on a common denominator (mean-square error in this example) are correlated. This complicates the problem further. In general, the greater the discrepancy between numerator and denominator degrees of freedom

and the smaller the denominator degrees of freedom, the greater the dependence of the tests. The Bonferroni tests are best in this situation, although Feingold and Korsog (1986) offer some useful alternatives.

Example 7

Separate Variance Hypothesis Tests

The data in the *MJ20* data file are from Milliken and Johnson (1984). They are the results of a paired-associate learning task. *GROUP* describes the type of drug administered; *LEARNING* is the amount of material learned during testing. First we perform Levene's test (Levene, 1960) to determine if the variances are equal across cells.

The input is:

```
ANOVA
USE MJ20
SAVE MJRESIDS / RESID DATA
DEPEND LEARNING
CATEGORY GROUP
ESTIMATE
USE MJRESIDS
LET RESIDUAL = ABS (RESIDUAL)
CATEGORY GROUP
DEPEND RESIDUAL
ESTIMATE
```

The following is the ANOVA table of the absolute residuals:

Dependent Variable	RESIDUAL					
N						29
Multiple R						0.675
Squared Multiple R						0.455
Analysis of Variance						
Source	Type III SS	df	Mean Squares	F-ratio	p-value	
GROUP	30.603	3	10.201	6.966	0.001	
Error	36.608	25	1.464			

Notice that the *F-ratio* is significant, indicating that the separate variances test is advisable. Let us do several single-degree-of-freedom tests, following Milliken and Johnson. The first is for comparing all drugs against the control; the second tests the hypothesis that groups 2 and 3 together are not significantly different from group 4.

The input is:

```
ANOVA
USE MJ20
CATEGORY GROUP
DEPEND LEARNING
ESTIMATE
HYPOTHESIS
SPECIFY 3*GROUP[1] = GROUP[2] +GROUP[3] + GROUP[4] / SEPARATE
TEST
HYPOTHESIS
SPECIFY 2*GROUP[4] = GROUP[2] +GROUP[3] / SEPARATE
TEST
```

The ANOVA table has been omitted because it is not valid when variances are unequal.

The output is:

Using separate variances estimate for error term.

A Matrix

1	2	3	4
0.000	4.000	0.000	0.000

Null Hypothesis Value for D

0.000

Null Hypothesis Contrast AB-D

-20.327

Contrast Estimate

Hypothesis	Estimate(AB-D)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	-20.327	4.776	-20.447	-20.208

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	242.720	1	242.720	18.115	0.004
Error	95.085	7.096	13.399		

Using separate variances estimate for error term.

A Matrix

1	2	3	4
0.000	-2.000	-3.000	-3.000

Null Hypothesis Value for D

0.000

Null Hypothesis Contrast AB-D

7.208

Contrast Estimate

Hypothesis	Estimate (AB-D)	Standard Error	95.0% Confidence Interval	
			Lower	Upper
A	7.208	1.679	7.166	7.250

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	65.634	1	65.634	18.431	0.000
Error	72.452	20.346	3.561		

Example 8

Analysis of Covariance

Winer, Brown and Michels (1991) uses the *COVAR* data file for an analysis of covariance in which *X* is the covariate and *TREAT* is the treatment. Cases do not need to be ordered by the grouping variable *TREAT*.

Before analyzing the data with an analysis of covariance model, be sure there is no significant interaction between the covariate and the treatment. The assumption of no interaction is often called the homogeneity of slopes assumption because it is tantamount to saying that the slope of the regression line of the dependent variable onto the covariate should be the same in all cells of the design.

Parallelism is easy to test with a preliminary model. Use GLM to estimate this model with the interaction between treatment (*TREAT*) and covariate (*X*) in the model.

The input is:

```
GLM
USE COVAR
CATEGORY TREAT
MODEL Y = CONSTANT + TREAT + X + TREAT*X
ESTIMATE
```

The output is:

Dependent Variable	Y
N	21
Multiple R	0.921
Squared Multiple R	0.849

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
TREAT	6.693	2	3.346	5.210	0.019
X	15.672	1	15.672	24.399	0.000
TREAT*X	0.667	2	0.334	0.519	0.605
Error	9.635	15	0.642		

The probability value for the treatment by covariate interaction is 0.605, so the assumption of homogeneity of slopes is plausible.

Now, fit the usual analysis of covariance model by specifying:

```
ANOVA
  USE COVAR
  PLENGTH MEDIUM
  CATEGORY TREAT
  DEPEND Y
  COVARIATE X
  ESTIMATE
```

For incomplete factorials and similar designs, you still must specify a model (using GLM) to do analysis of covariance.

The output is:

```
Dependent Variable : Y
N : 21
Multiple R : 0.916
Squared Multiple R : 0.839
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
TREAT	16.932	2	8.466	13.970	0.000
X	16.555	1	16.555	27.319	0.000
Error	10.302	17	0.606		

Least Squares Means

Factor	Level	LS Mean	Standard Error	N
TREAT	1	4.888	0.307	7.000
TREAT	2	7.076	0.309	7.000
TREAT	3	6.750	0.294	7.000

* Means are computed after adjusting covariate effect.

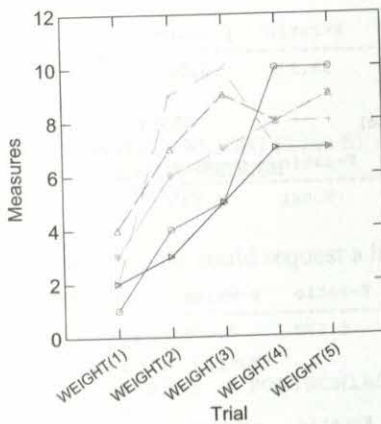
The treatment adjusted for the covariate is significant. There is a significant difference among the three treatment groups. Also, notice that the coefficient for the covariate is significant ($F\text{-ratio} = 27.319$, $p\text{-value} < 0.0005$). If it were not, the analysis of covariance could be taking away a degree of freedom without reducing mean-square error enough to help you.

SYSTAT computes the adjusted cell means the same way it computes estimates when saving residuals. Model terms (main effects and interactions) that do not contain categorical variables (covariates) are incorporated into the equation by adding the product of the coefficient and the mean of the term for computing estimates. The grand mean (*CONSTANT*) is included in computing the estimates.

Example 9

One-Way Repeated Measures

In this example, six rats were weighed at the end of each of five weeks. A plot of each rat's weight over the duration of the experiment is shown below:



ANOVA is the simplest way to analyze this one-way model. Because we have no categorical variable(s), SYSTAT generates only the constant (grand mean) in the model. To obtain individual single-degree-of-freedom orthogonal polynomials, the input is:

```
ANOVA
  USE RATS
  DEPEND WEIGHT(1..5) / REPEAT NAME="Time"
  PLENGTH MEDIUM
  ESTIMATE
```

The output is:

N of Cases Processed : 6

Dependent Variable Means

WEIGHT (1)	WEIGHT (2)	WEIGHT (3)	WEIGHT (4)	WEIGHT (5)
2.500	5.833	7.167	8.000	8.333

Univariate and Multivariate Repeated Measures Analysis

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
Time	134.467	4	33.617	16.033	0.000	0.004	0.002
Error	41.933	20	2.097				
Greenhouse-Geisser Epsilon			0.342				
Huynh-Feldt Epsilon			0.427				

Single Degree of Freedom Polynomial Contrasts**Polynomial Test of Order 1 (Linear)**

Source	SS	df	Mean Squares	F-ratio	p-value
Time	114.817	1	114.817	38.572	0.002
Error	14.883	5	2.977		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	18.107	1	18.107	7.061	0.045
Error	12.821	5	2.564		

Polynomial Test of Order 3 (Cubic)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	1.350	1	1.350	0.678	0.448
Error	9.950	5	1.990		

Polynomial Test of Order 4

Source	SS	df	Mean Squares	F-ratio	p-value
Time	0.193	1	0.193	0.225	0.655
Error	4.279	5	0.856		

Multivariate Repeated Measures Analysis**Test of: Time**

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.011	4	2	43.007	0.023
Pillai Trace	0.989	4	2	43.007	0.023
Hotelling-Lawley Trace	86.014	4	2	43.007	0.023

The Huynh-Feldt *p*-value (0.002) does not differ from the *p*-value for the *F*-ratio to any significant degree. Compound symmetry appears to be satisfied and weight changes significantly over the five trials.

The polynomial tests indicate that most of the trials effect can be accounted for by a linear trend across time. In fact, the sum of squares for *TIME* is 134.467, and the sum of squares for the linear trend is almost as large (114.817). Thus, the linear polynomial accounts for roughly 85% of the change across the repeated measures.

Unevenly Spaced Polynomials

Sometimes the underlying metric of the profiles is not evenly spaced. Let's assume that the fifth weight was measured after the tenth week instead of the fifth. In that case, the default polynomials have to be adjusted for the uneven spacing. These adjustments do not affect the overall repeated measures tests of each effect (univariate or multivariate), but they partition the sum of squares differently for the single-degree-of-freedom tests.

The input is:

```
ANOVA
  USE RATS
  DEPEND WEIGHT(1.. 5) / REPEAT=5(1 2 3 4 10) NAME="Time"
  PLENGTH MEDIUM
  ESTIMATE
```

Alternatively, you could request a hypothesis test, specifying the metric for the polynomials:

```
HYPOTHESIS
  WITHIN 'Time'
  CONTRAST / POLYNOMIAL METRIC=1,2,3,4,10
  TEST
```

The last point has been spread out further to the right.

The output is:

Univariate and Multivariate Repeated Measures Analysis

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
Time	134.467	4	33.617	16.033	0.000	0.004	0.002
Error	41.933	20	2.097				
Greenhouse-Geisser Epsilon			0.342				
Huynh-Feldt Epsilon			0.427				

Single Degree of Freedom Polynomial Contrasts

Polynomial Test of Order 1 (Linear)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	67.213	1	67.213	23.959	0.004
Error	14.027	5	2.805		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	62.283	1	62.283	107.867	0.000
Error	2.887	5	0.577		

The significance tests for the linear and quadratic trends differ from those for the evenly spaced polynomials. Before, the linear trend was strongest; now, the quadratic polynomial has the most significant results ($F\text{-ratio} = 107.9$, $p\text{-value} < 0.0005$).

You may have noticed that although the univariate F-tests for the polynomials are different, the multivariate test is unchanged. The latter measures variation across all components. The ANOVA table for the combined components is not affected by the metric of the polynomials.

Difference Contrasts

If you do not want to use polynomials, you can specify a **C** matrix that contrasts adjacent weeks. After estimating the model, use the following input:

```
HYPOTHESIS
WITHIN 'Time'
CONTRAST / ADJDIFF
TEST
```

The output is:

Multivariate Repeated Measures Analysis

Test of: Time

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.011	4	2	43.007	0.023
Pillai Trace	0.989	4	2	43.007	0.023
Hotelling-Lawley Trace	86.014	4	2	43.007	0.023

Notice the **C** matrix that this command generates. In this case, each of the univariate F-tests covers the significance of the difference between the adjacent weeks indexed by the **C** matrix. For example, the $F\text{-ratio} = 17.241$ shows that the first and second weeks differ significantly. The third and fourth weeks do not differ ($F\text{-ratio} = 0.566$). Unlike polynomials, these contrasts are not orthogonal.

Summing Effects

To sum across weeks, the input is:

```
HYPOTHESIS
WITHIN 'Time'
CONTRAST / SUM
TEST
```

The output is:

C Matrix

	1	2	3	4	5
1	1.000	-1.000	0.000	0.000	0.000
2	0.000	1.000	-1.000	0.000	0.000
3	0.000	0.000	1.000	-1.000	0.000
4	0.000	0.000	0.000	1.000	-1.000

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
1	66.667	1	66.667	17.241	0.009
Error	19.333	5	3.867		
2	10.667	1	10.667	40.000	0.001
Error	1.333	5	0.267		
3	4.167	1	4.167	0.566	0.486
Error	36.833	5	7.367		
4	0.667	1	0.667	2.500	0.175
Error	1.333	5	0.267		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.011	43.007	4, 2	0.023
Pillai Trace	0.989	43.007	4, 2	0.023
Hotelling-Lawley Trace	86.014	43.007	4, 2	0.023

In this example, you are testing whether the overall weight (across weeks) significantly differs from 0. Naturally, the *F-ratio* is significant. Notice the *C* matrix that is generated. It is simply a set of 1's that, in the equation $BC' = 0$, sum all the coefficients in *B*. In a group-by-trials design, this *C* matrix is useful for pooling trials and analyzing group effects.

Custom Contrasts

To test any arbitrary contrast effects between dependent variables, you can use the *C* matrix, which has the same form (without a column for the *CONSTANT*) as the *A* matrix. The following commands test a linear trend across the five trials:

```
HYPOTHESIS
CMATRIX [-2 -1 0 1 2]
TEST
```

The output is:

C Matrix

1	2	3	4	5
1.000	1.000	1.000	1.000	1.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	6080.167	1	6080.167	295.632	0.000
Error	102.833	5	20.567		

C Matrix

1	2	3	4	5
-2.000	-1.000	0.000	1.000	2.000

Test of Hypothesis

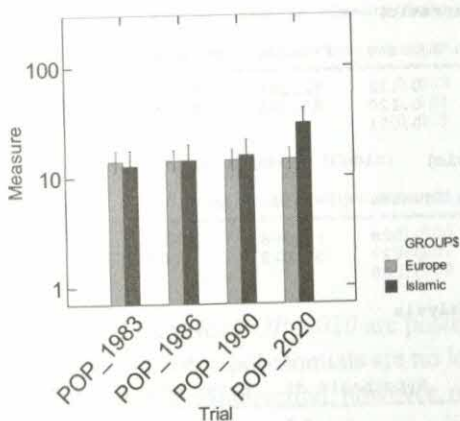
Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	1148.167	1	1148.167	38.572	0.002
Error	148.833	5	29.767		

Example 10

Repeated Measures ANOVA for One Grouping Factor and One Within Factor with Ordered Levels

The following example uses estimates of population for 1983, 1986, and 1990 and projections for 2020 for 57 countries from the *OURWORLD* data file. The data are log transformed before analysis. Here you compare trends in population growth for European and Islamic countries. The variable *GROUP\$* contains codes for these groups plus a third code for New World countries (we exclude these countries from this analysis). To create a bar chart of the data after using *YLOG* to log transform them:

```
USE OURWORLD
SELECT GROUP$ <> 'NewWorld'
BAR pop_1983.. pop_2020 / REPEAT OVERLAY YLOG,
GROUP=group$SERROR FILL=.35,.8
```



To perform a repeated measures analysis:

ANOVA

USE OURWORLD

SELECT GROUP\$ <> 'NewWorld'

CATEGORY GROUP\$

LET (POP_1983, POP_1986, POP_1990, POP_2020) = L10(@)

DEPEND POP_1983 POP_1986 POP_1990 POP_2020 / REPEAT=4 NAME='Time'

PLENGTH MEDIUM

ESTIMATE

The output is:

Univariate and Multivariate Repeated Measures Analysis

Between Subjects

Source	SS	df	Mean Squares	F-ratio	p-value
GROUP\$	0.233	1	0.233	0.257	0.616
Error	30.794	34	0.906		

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
Time	0.835	3	0.278	235.533	0.000	0.000	0.000
Time*GROUP\$	0.739	3	0.246	208.352	0.000	0.000	0.000
Error	0.121	102	0.001				

Greenhouse-Geisser Epsilon : 0.528
Huynh-Feldt Epsilon : 0.566

Single Degree of Freedom Polynomial Contrasts

Polynomial Test of Order 1 (Linear)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	0.675	1	0.675	370.761	0.000
Time*GROUP\$	0.583	1	0.583	320.488	0.000
Error	0.062	34	0.002		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	0.132	1	0.132	92.246	0.000
Time*GROUP\$	0.128	1	0.128	89.095	0.000
Error	0.049	34	0.001		

Polynomial Test of Order 3 (Cubic)

Source	SS	df	Mean Squares	F-ratio	p-value
Time	0.028	1	0.028	96.008	0.000
Time*GROUP\$	0.027	1	0.027	94.828	0.000
Error	0.010	34	0.000		

Multivariate Repeated Measures Analysis

Test of: Time

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.063	3	32	157.665	0.000
Pillai Trace	0.937	3	32	157.665	0.000
Hottelling-Lawley Trace	14.781	3	32	157.665	0.000

Test of: Time*GROUP\$

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.076	3	32	130.336	0.000
Pillai Trace	0.924	3	32	130.336	0.000
Hottelling-Lawley Trace	12.219	3	32	130.336	0.000

The within-subjects results indicate highly significant linear, quadratic, and cubic changes across time. The pattern of change across time for the two groups also differs significantly (that is, the *TIME * GROUP\$* interactions are highly significant for all three tests).

Notice that there is a larger gap in time between 1990 and 2020 than between the other values. Let's incorporate "real time" in the analysis with the following specification:

```
DEPEND POP_1983 POP_1986 POP_1990 POP_2020/REPEAT=4 (83,86,90,120),
NAME='TIME'
ESTIMATE
```

The results for the orthogonal polynomials are shown below:

Single Degree of Freedom Polynomial Contrasts

Polynomial Test of Order 1 (Linear)

Source	SS	df	Mean Squares	F-ratio	p-value
TIME	0.831	1	0.831	317.273	0.000
TIME*GROUP\$	0.737	1	0.737	281.304	0.000
Error	0.089	34	0.003		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
TIME	0.003	1	0.003	4.402	0.043
TIME*GROUP\$	0.001	1	0.001	1.562	0.220
Error	0.025	34	0.001		

Polynomial Test of Order 3 (Cubic)

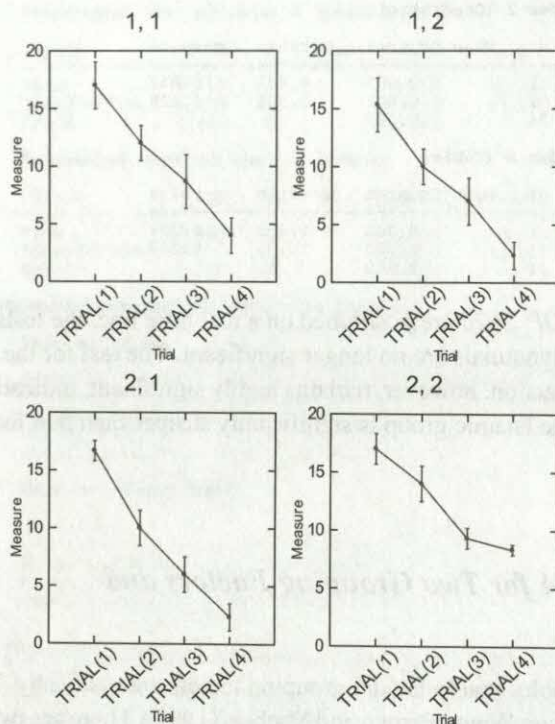
Source	SS	df	Mean Squares	F-ratio	p-value
TIME	0.000	1	0.000	1.653	0.207
TIME*GROUP\$	0.000	1	0.000	1.733	0.197
Error	0.006	34	0.000		

When the values for *POP_2020* are positioned on a real time line, the tests for quadratic and cubic polynomials are no longer significant. The test for the linear *TIME * GROUP\$* interaction, however, remains highly significant, indicating that the slope across time for the Islamic group is significantly steeper than that for the European countries.

Example 11

Repeated Measures ANOVA for Two Grouping Factors and One Within Factor

Repeated measures enables you to handle grouping factors automatically. The following example is from Winer, Brown and Michels (1991). There are two grouping factors (*ANXIETY* and *TENSION*) and one trial factor in the file *REPEAT1*. The following is a dot display of the average responses across trials for each of the four combinations of *ANXIETY* and *TENSION*.



The input is:

```
ANOVA
  USE REPEAT1
  LET TENS = TENSION
  DOT TRIAL(1..4) / Group=anxiety,tens, LINE,REPEAT,ERROR
  CATEGORY ANXIETY TENSION
  DEPEND TRIAL(1 .. 4) / REPEAT NAME='Trial'
  PLENGTH MEDIUM
  ESTIMATE
```

The model also includes an interaction between the grouping factors (*ANXIETY * TENSION*).

The output is:

Univariate and Multivariate Repeated Measures Analysis

Between Subjects

Source	SS	df	Mean Squares	F-ratio	p-value
ANXIETY	10.083	1	10.083	0.978	0.352
TENSION	8.333	1	8.333	0.808	0.395
ANXIETY*TENSION	80.083	1	80.083	7.766	0.024
Error	82.500	8	10.313		

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
Trial	991.500	3	330.500	152.051	0.000	0.000	0.000
Trial*ANXIETY	8.417	3	2.806	1.291	0.300	0.300	0.301
Trial*TENSION	12.167	3	4.056	1.866	0.162	0.197	0.169
Trial*ANXIETY*TENSION	12.750	3	4.250	1.955	0.148	0.185	0.155
Error	52.167	24	2.174				

Greenhouse-Geisser Epsilon | 0.536
Huynh-Feldt Epsilon | 0.902

Single Degree of Freedom Polynomial Contrasts

Polynomial Test of Order 1 (Linear)

Source	SS	df	Mean Squares	F-ratio	p-value
Trial	984.150	1	984.150	247.845	0.000
Trial*ANXIETY	1.667	1	1.667	0.420	0.535
Trial*TENSION	10.417	1	10.417	2.623	0.144
Trial*ANXIETY*TENSION	9.600	1	9.600	2.418	0.159
Error	31.767	8	3.971		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
Trial	6.750	1	6.750	3.411	0.102
Trial*ANXIETY	3.000	1	3.000	1.516	0.253
Trial*TENSION	0.083	1	0.083	0.042	0.843
Trial*ANXIETY*TENSION	0.333	1	0.333	0.168	0.692
Error	15.833	8	1.979		

Polynomial Test of Order 3 (Cubic)

Source	SS	df	Mean Squares	F-ratio	p-value
Trial	0.600	1	0.600	1.051	0.335
Trial*ANXIETY	3.750	1	3.750	6.569	0.033
Trial*TENSION	1.667	1	1.667	2.920	0.126
Trial*ANXIETY*TENSION	2.817	1	2.817	4.934	0.057
Error	4.567	8	0.571		

Multivariate Repeated Measures Analysis

Test of: Trial

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.015	3	6	127.686	0.000
Pillai Trace	0.985	3	6	127.686	0.000
Hotelling-Lawley Trace	63.843	3	6	127.686	0.000

Test of: Trial*ANXIETY

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.244	3	6	6.183	0.029
Pillai Trace	0.756	3	6	6.183	0.029
Hotelling-Lawley Trace	3.091	3	6	6.183	0.029

Test of: Trial*TENSION

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.361	3	6	3.546	0.088
Pillai Trace	0.639	3	6	3.546	0.088
Hotelling-Lawley Trace	1.773	3	6	3.546	0.088

Test of: Trial*ANXIETY*TENSION

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.328	3	6	4.099	0.067
Pillai Trace	0.672	3	6	4.099	0.067
Hotelling-Lawley Trace	2.050	3	6	4.099	0.067

In the within-subjects table, you see that the trial effect is highly significant ($F\text{-ratio} = 152.1$, $p\text{-value} < 0.0005$). Below that table, we see that the linear trend across trials (*Polynomial Order 1*) is highly significant ($F\text{-ratio} = 247.8$, $p\text{-value} < 0.0005$). The hypothesis sum of squares for the linear, quadratic, and cubic polynomials sum to the total hypothesis sum of squares for trials (that is, $984.15 + 6.75 + 0.60 = 991.5$). Notice that the total sum of squares is 991.5, while that for the linear trend is 984.15. This means that the linear trend accounts for more than 99% of the variability across the four trials. The assumption of compound symmetry is not required for the test of linear trend—so you can report that there is a highly significant linear decrease across the four trials ($F\text{-ratio} = 247.8$, $p\text{-value} < 0.0005$).

Example 12**Repeated Measures ANOVA for Two Trial Factors**

Repeated Measures enables you to handle several trial factors, so we include an example with two trial factors. It is an experiment from Winer, Brown and Michels (1991), which has one grouping factor (*NOISE*) and two trials factors (*PERIODS* and *DIALS*). The trial factors must be sorted into a set of dependent variables (one for each pairing of the two factors groups). It is useful to label the levels with a convenient mnemonic. The file is set up with variables *P1D1* through *P3D3*. Variable *P1D2* indicates a score in the *PERIODS* = 1, *DIALS* = 2 cell. The data are in the file *REPEAT2*.

The input is:

```
ANOVA
  USE REPEAT2
  CATEGORY NOISE
  DEPEND P1D1 .. P3D3 / REPEAT=3,3 NAMES='period','dial'
  LENGTH MEDIUM
  ESTIMATE
```

Notice that REPEAT specifies that the two trial factors have three levels each. ANOVA assumes the subscript of the first factor will vary the slowest in the ordering of the dependent variables. If you have two repeated factors (*DAY* with four levels and *AMPM* with two levels), you should select eight dependent variables and type Repeat=4, 2. The repeated measures are selected in the following order:

```
DAY1_AM DAY1_PM DAY2_AM DAY2_PM DAY3_AM DAY3_PM DAY4_AM
DAY4_PM
```

From this indexing, it generates the proper main effects and interactions. When more than one trial factor is present, ANOVA lists each dependent variable and the associated level on each factor.

The output is:

Dependent Variable Means

P1D1	P1D2	P1D3	P2D1	P2D2
48.000	52.000	63.000	37.167	42.167

Dependent Variable Means

P2D3	P3D1	P3D2	P3D3
54.167	27.000	32.500	42.500

Univariate and Multivariate Repeated Measures Analysis

Between Subjects

Source	SS	df	Mean Squares	F-ratio	p-value
NOISE	468.167	1	468.167	0.752	0.435
Error	2491.111	4	622.778		

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
period	3722.333	2	1861.167	63.389	0.000	0.000	0.000
period*NOISE	333.000	2	166.500	5.671	0.029	0.057	0.029
Error	234.889	8	29.361				

Greenhouse-Geisser Epsilon	0.648
Huynh-Feldt Epsilon	1.000

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
dial	2370.333	2	1185.167	89.823	0.000	0.000	0.000
dial*NOISE	50.333	2	25.167	1.907	0.210	0.215	0.210
Error	105.556	8	13.194				

Greenhouse-Geisser Epsilon : 0.917

Huynh-Feldt Epsilon : 1.000

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
period*dial	10.667	4	2.667	0.336	0.850	0.729	0.850
period*dial*NOISE	11.333	4	2.833	0.357	0.836	0.716	0.836
Error	127.111	16	7.944				

Greenhouse-Geisser Epsilon : 0.513

Huynh-Feldt Epsilon : 1.000

Single Degree of Freedom Polynomial Contrasts**Polynomial Test of Order 1 (Linear)**

Source	SS	df	Mean Squares	F-ratio	p-value
period	3721.000	1	3721.000	73.441	0.001
period*NOISE	225.000	1	225.000	4.441	0.103
Error	202.667	4	50.667		
dial	2256.250	1	2256.250	241.741	0.000
dial*NOISE	6.250	1	6.250	0.670	0.459
Error	37.333	4	9.333		
period*dial	0.375	1	0.375	0.045	0.842
period*dial*NOISE	1.042	1	1.042	0.125	0.742
Error	33.333	4	8.333		

Polynomial Test of Order 2 (Quadratic)

Source	SS	df	Mean Squares	F-ratio	p-value
period	1.333	1	1.333	0.166	0.705
period*NOISE	108.000	1	108.000	13.407	0.022
Error	32.222	4	8.056		
dial	114.083	1	114.083	6.689	0.061
dial*NOISE	44.083	1	44.083	2.585	0.183
Error	68.222	4	17.056		
period*dial	3.125	1	3.125	0.815	0.418
period*dial*NOISE	0.125	1	0.125	0.033	0.865
Error	15.333	4	3.833		

Polynomial Test of Order 3 (Cubic)

Source	SS	df	Mean Squares	F-ratio	p-value
period*dial	6.125	1	6.125	0.750	0.435
period*dial*NOISE	3.125	1	3.125	0.383	0.570
Error	32.667	4	8.167		

Polynomial Test of Order 4

Source	SS	df	Mean Squares	F-ratio	p-value
period*dial	1.042	1	1.042	0.091	0.778
period*dial*NOISE	7.042	1	7.042	0.615	0.477
Error	45.778	4	11.444		

Multivariate Repeated Measures Analysis

Test of: period

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.051	2	3	28.145	0.011
Pillai Trace	0.949	2	3	28.145	0.011
Hotelling-Lawley Trace	18.764	2	3	28.145	0.011

Test of: period*NOISE

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.156	2	3	8.111	0.062
Pillai Trace	0.844	2	3	8.111	0.062
Hotelling-Lawley Trace	5.407	2	3	8.111	0.062

Test of: dial

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.016	2	3	91.456	0.002
Pillai Trace	0.984	2	3	91.456	0.002
Hotelling-Lawley Trace	60.971	2	3	91.456	0.002

Test of: dial*NOISE

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.565	2	3	1.155	0.425
Pillai Trace	0.435	2	3	1.155	0.425
Hotelling-Lawley Trace	0.770	2	3	1.155	0.425

Test of: period*dial

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.001	4	1	331.445	0.041
Pillai Trace	0.999	4	1	331.445	0.041
Hotelling-Lawley Trace	1325.780	4	1	331.445	0.041

Test of: period*dial*NOISE

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.000	4	1	581.875	0.031
Pillai Trace	1.000	4	1	581.875	0.031
Hotelling-Lawley Trace	2327.500	4	1	581.875	0.031

The input is:

GLM

USE REPEAT2

CATEGORY NOISE

MODEL P1D1 .. P3D3 = CONSTANT + NOISE / REPEAT=3,3,
NAMES='period','dial'

PLENGTH MEDIUM

ESTIMATE

Example 13

Repeated Measures Analysis of Covariance

To do repeated measures analysis of covariance, where the covariate varies within subjects, you would have to set up your model like a split plot with a different record for each measurement.

This example is from Winer, Brown and Michels (1991). This design has two trials (*DAY1* and *DAY2*), one covariate (*AGE*), and one grouping factor (*SEX*). The data are in the file *WINER*.

The input is:

```
ANOVA
  USE WINER
  CATEGORY SEX
  DEPEND DAY(1 .. 2) / REPEAT NAME='day'
  COVARIATE AGE
  ESTIMATE
```

The output is:

Dependent Variable Means

DAY(1)	DAY(2)
16.500	11.875

Univariate Repeated Measures Analysis

Between Subjects

Source	SS	df	Mean Squares	F-ratio	p-value
SEX	44.492	1	44.492	3.629	0.115
AGE	166.577	1	166.577	13.587	0.014
Error	61.298	5	12.260		

Within Subjects

Source	SS	df	Mean Squares	F-ratio	p-value	G-G	H-F
day	22.366	1	22.366	17.899	0.008	.	.
day*SEX	0.494	1	0.494	0.395	0.557	.	.
day*AGE	0.127	1	0.127	0.102	0.763	.	.
Error	6.248	5	1.250				

Greenhouse-Geisser Epsilon | .
Huynh-Feldt Epsilon | .

The *F-ratio* for the covariate and its interactions, namely *AGE* (13.587) and *DAY * AGE* (0.102), are not ordinarily published; however, they help you understand the adjustment made by the covariate.

This analysis did not test the homogeneity of slopes assumption. If you want to test the homogeneity of slopes assumption, run the following model in GLM first:

```
MODEL day(1 .. 2) = CONSTANT + sex + age + sex*age / REPEAT
```

Then check to see if the *SEX * AGE* interaction is significant.

To use GLM:

```
GLM
  USE WINER
  CATEGORY SEX
  MODEL DAY(1 .. 2) = CONSTANT + SEX + AGE / REPEAT NAME='day'
  ESTIMATE
```

Computation

Algorithms

Centered sum of squares and cross-products are accumulated using provisional algorithms. Linear systems, including those involved in hypothesis testing, are solved by using forward and reverse sweeping (Dempster, 1969). Eigensystems are solved with Householder tridiagonalization and implicit QL iterations. For further information, see Wilkinson and Reinsch (1971) or Chambers (1977).

References

- Afifi, A. A. and Azen, S. P. (1972). *Statistical analysis: A computer-oriented approach*. New York: Academic Press.
- Bartlett, M. S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society*, Series B, 9, 176–197.
- * Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Burnham, K. P., and Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.
- Chambers, J. M. (1977). *Computational methods for data analysis*. New York: John Wiley & Sons.

- Cochran, W. G., and Cox, G. M. (1957). *Experimental designs*, 2nd ed. New York: John Wiley & Sons.
- Daniel, C. (1960). Locating outliers in factorial experiments. *Technometrics*, 2, 149–156.
- Dempster, A.P. (1969). *Elements of continuous multivariate analysis*. San Francisco: Addison-Wesley.
- Feingold, M. and Korsog, P. E. (1986). The correlation and dependence between two f statistics with the same denominator. *The American Statistician*, 40, 218–220.
- * Hurvich, C.M., and Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.
- John, P. W. M. (1971). *Statistical design and analysis of experiments*. New York: MacMillan.
- Kutner, M. H. (1974). Hypothesis testing in linear models (Eisenhart Model I). *The American Statistician*, 28, 98–100.
- Levene, H. (1960). Robust tests for equality of variance. I. Olkin, ed., *Contributions to Probability and Statistics*. Palo Alto, Calif.: Stanford University Press, 278–292.
- Miller, R. (1985). Multiple comparisons. Kotz, S. and Johnson, N. L., eds., *Encyclopedia of Statistical Sciences*, vol. 5. New York: John Wiley & Sons, 679–689.
- Milliken, G. A. and Johnson, D. E. (1984). Analysis of messy data, Vol. 1: *Designed Experiments*. New York: Van Nostrand Reinhold Company.
- Morrison, D. F. (2004). *Multivariate statistical methods*, 4th ed. Pacific Grove, CA: Duxbury Press.
- Kutner, M.H, Nachtsheim, C.J., Neter, J., and Li, W. (2004). *Applied linear statistical models*, 5th ed. Irwin: McGraw-Hill.
- * Pillai, K. C. S. (1960). *Statistical table for tests of multivariate hypotheses*. Manila: The Statistical Center, University of Phillipines.
- * Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons.
- * Schatzoff, M. (1966). Exact distributions of Wilk's likelihood ratio criterion. *Biometrika*, 53, 347–358.
- Scheffé, H. (1959). *The analysis of variance*. New York: John Wiley & Sons.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.
- * Searle, S. R. (1971). *Linear models*. New York: John Wiley & Sons.
- Speed, F. M., Hocking, R. R., and Hackney, O. P. (1978). Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association*, 73, 105–112.
- * Timm, N.H. (2002). *Applied multivariate analysis*. New York: Springer-Verlag.
- Wilkinson, L. (1975). Response variable hypotheses in the multivariate analysis of variance. *Psychological Bulletin*, 82, 408–412.
- * Wilkinson, L. (1977). Confirmatory rotation of MANOVA canonical variates. *Multivariate*

- Behavioral Research*, 12, 487–494.
- Wilkinson, J.H. and Reinsch, C. (Eds.). (1971). *Linear Algebra*, Vol. 2, *Handbook for automatic computation*. New York: Springer-Verlag.
- Winer, B. J., Brown, D. R., and Michels, K. M. (1991). *Statistical principles in experimental design*, 3rd ed. New York: McGraw-Hill.

(* indicates additional reference.)

Linear Models III: General Linear Models

Leland Wilkinson and Mark Coward

General Linear Model (GLM) can estimate and test any univariate or multivariate general linear model, including those for multiple regression, analysis of variance or covariance, and other procedures such as discriminant analysis and principal components. With the general linear model, you can explore randomized block designs, incomplete block designs, fractional factorial designs, Latin square designs, split plot designs, crossover designs, nesting, and more. The model is:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{e}$$

where \mathbf{Y} is a vector or matrix of dependent variables, \mathbf{X} is a vector or matrix of independent variables, \mathbf{B} is a vector or matrix of regression coefficients, and \mathbf{e} is a vector or matrix of random errors. See Searle (1971), Winer, Brown and Michels (1991), Kutner et al. (2004), or Cohen (2002) for details.

Moreover, GLM also features the means model for missing cells designs. Widely favored for this purpose by statisticians (Hocking, 1985; Milliken and Johnson, 1984; Searle, 1987), the means model allows tests of hypothesis in missing cells designs (using what are often called Type IV sum of squares). Furthermore, the means model allows direct tests of simple hypotheses (for example, within levels of other factors). Finally, the means model allows easier use of population weights to reflect differences in subclass sizes.

The GLM module provides fifteen tests for pairwise comparisons based on the structure of data and the error rate to be controlled. The pairwise comparison tests are commonly named as post hoc tests; here tests are determined based on the assumptions on variance, viz., equal or unequal variances. One can use post hoc tests after fitting the model to check the differences between pairs of means.

In multivariate models, **Y** is a matrix of continuous measures. The **X** matrix can be either continuous or categorical dummy variables, according to the type of model. For discriminant analysis, **X** is a matrix of dummy variables, as in analysis of variance. For principal components analysis, **X** is constant (a single column of 1's). For canonical correlation, **X** is usually a matrix of continuous right-hand variables (and **Y** is the matrix of left-hand variables).

For some multivariate models, it may be easier to use ANOVA, which can handle models with multiple dependent variables and zero, one, or more categorical independent variables (that is, only the constant is present in the former). ANOVA automatically generates interaction terms for the design factor.

SYSTAT offers three tests for checking normality: Kolmogorov-Smirnov (Lilliefors), Anderson-Darling, and Shapiro-Wilk test; and Levene's test for checking the homogeneity of variances. You can select any of the three types of sum of squares: Type I, Type II and Type III, for the analysis.

After the parameters of a model have been estimated, they can be tested by any general linear hypothesis of the following form:

$$ABC' = D$$

where **A** is a matrix of linear weights on coefficients across the independent variables (the rows of **B**), **C** is a matrix of linear weights on the coefficients across dependent variables (the columns of **B**), **B** is the matrix of regression coefficients or effects, and **D** is a null hypothesis matrix (usually a null matrix).

For the multivariate models described in this chapter, by default the **C** matrix is an identity matrix, and the **D** matrix is null. The **A** matrix can have several different forms, but these are all submatrices of an identity matrix and are easily formed.

The **A** matrix, **C** matrix, and **D** matrix are available for hypothesis testing in multivariate models. You can test parameters of the multivariate model estimated or factor the quadratic form of your model into orthogonal components.

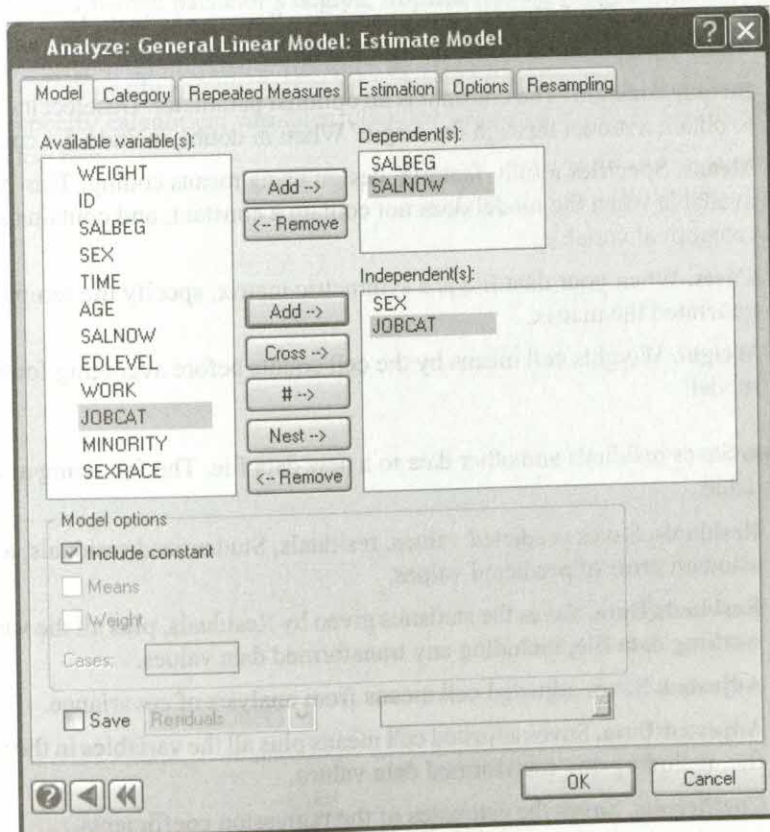
Resampling procedures are available in this feature.

General Linear Models in SYSTAT

Model Estimation (in GLM)

To specify a general linear model using GLM, from the menus choose:

Analyze
General Linear Model (GLM)
Estimate Model...



You can specify any multivariate linear model with General Linear Model. You must select the variables to include in the desired model.

Dependent(s). The variable(s) you want to examine. The dependent variable(s) should be continuous numeric variables (for example, *INCOME*).

Independent(s). Select one or more continuous or categorical variables (grouping variables). Independent variables that are not denoted as categorical are considered covariates. Unlike ANOVA, GLM does not automatically include and test all interactions. With GLM, you have to build your model. Suppose you want interactions or nested variables in your model, you need to build these components using the Cross and Nest buttons. To include lower-order effects with the interaction term, use the # button; e.g., $A \# B = A + B + A * B$.

Model options. The following model options allow you to include a constant in your model, do a means model, specify the sample size, and weight cell means.

- **Include constant.** The constant is an optional parameter. Deselect Include constant to obtain a model through the origin. When in doubt, include the constant.
- **Means.** Specifies a fully factorial design using means coding. This option is available when the model does not contain a constant, and contains at least one categorical variable.
- **Cases.** When your data file is a symmetric matrix, specify the sample size that generated the matrix.
- **Weight.** Weights cell means by the cell counts before averaging for the Means model.

Save. Saves residuals and other data to a new data file. The following alternatives are available:

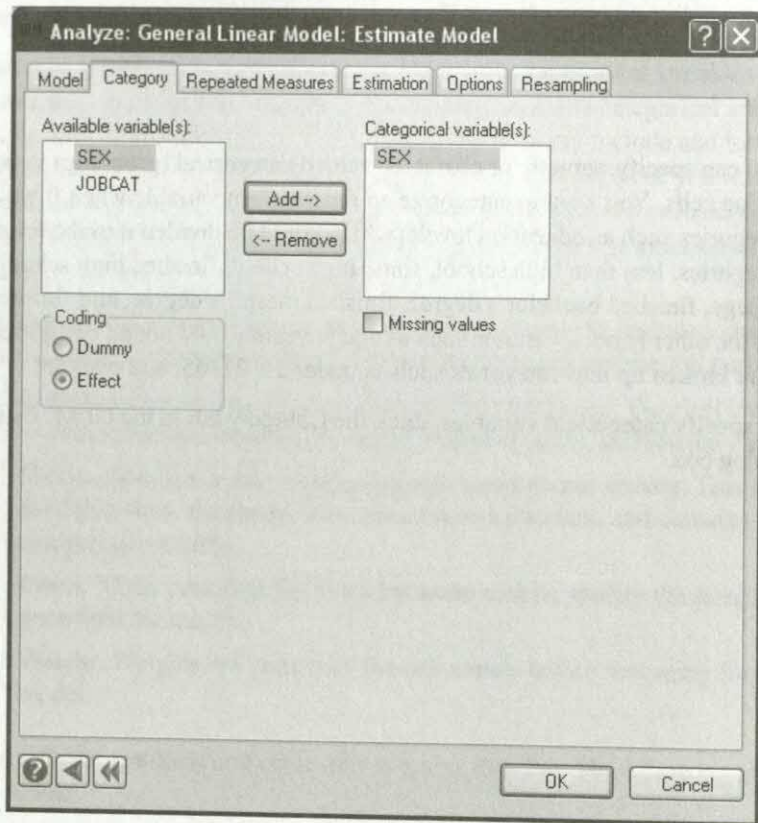
- **Residuals.** Saves predicted values, residuals, Studentized residuals, and the standard error of predicted values.
- **Residuals/Data.** Saves the statistics given by Residuals, plus all the variables in the working data file, including any transformed data values.
- **Adjusted.** Saves adjusted cell means from analysis of covariance.
- **Adjusted/Data.** Saves adjusted cell means plus all the variables in the working data file, including any transformed data values.
- **Coefficients.** Saves the estimates of the regression coefficients.
- **Model.** Saves statistics given in Residuals and the variables used in the model.
- **Partial.** Saves partial residuals for univariate model.

- **Partial/Data.** Saves partial residuals plus all the variables in the working data file, including any transformed data values.

Category

You can specify numeric or character-valued categorical (grouping) variables that define cells. You want to categorize an independent variable when it has several categories such as education levels, which could be divided into the following categories: less than high school, some high school, finished high school, some college, finished bachelor's degree, finished master's degree, and finished doctorate. On the other hand, a variable such as age in years would not be categorical unless age were broken up into categories such as under 21, 21–65, and over 65.

To specify categorical variables, click the Category tab in the GLM: Estimate Model dialog box.



Missing values. Specifies the cases with missing values for the categorical variable(s) to be included as a separate category in the analysis.

Coding. You can select to use one of two different coding methods:

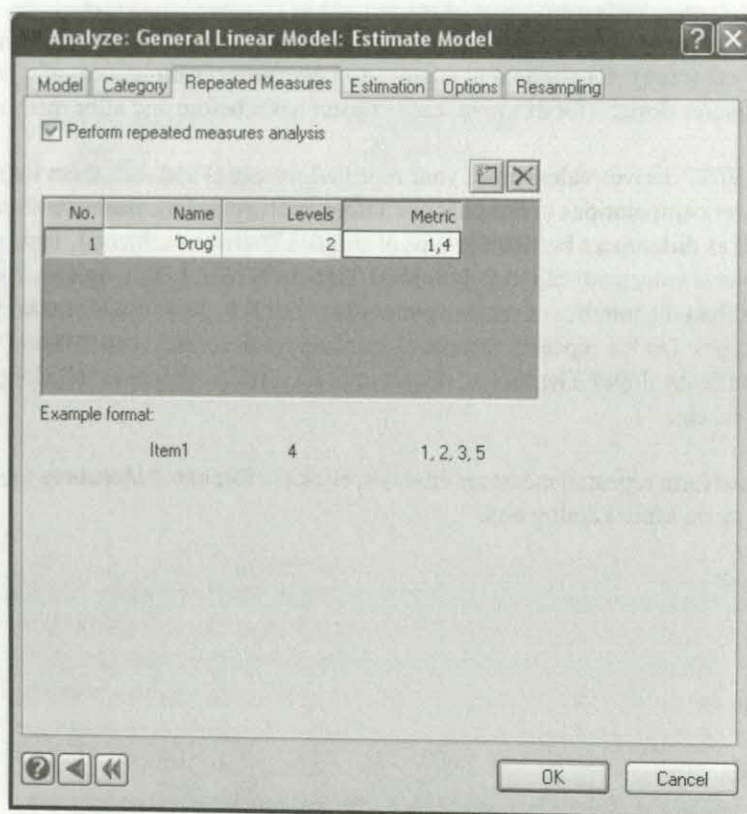
- **Dummy.** Produces dummy codes for the design variables instead of effect codes. Coding of dummy variables is the classic analysis of variance parameterization, in which the sum of effects estimated for a classifying variable is 0. If your categorical variable has k categories, $k - 1$ dummy variables are created.
- **Effect.** Produces parameter estimates that are differences from group means.

Repeated Measures

In a repeated measures design, the same variable is measured several times for each subject (case). A paired-comparison t-test is the most simple form of a repeated measures design (for example, each subject has a before and after measure).

SYSTAT derives values from your repeated measures and uses them in general linear model computations to test changes across the repeated measures (within subjects) as well as differences between groups of subjects (between subjects). Tests of the within-subjects values are called **Polynomial Tests of Order 1, 2,...**, up to k , where k is one less than the number of repeated measures. The first polynomial is used to test linear changes: Do the repeated responses increase (or decrease) around a line with a significant slope? The second polynomial tests if the responses fall along a quadratic curve, etc.

To perform repeated measures analysis, click the Repeated Measures tab in the GLM: Estimate Model dialog box.



Suppose you select Perform repeated measures analysis, SYSTAT treats the dependent variables as a set of repeated measures. Optionally, you can assign a name for each set of repeated measures, specify the number of levels, and specify the metric for unevenly spaced repeated measures.

Name. Name that identifies each set of repeated measures.

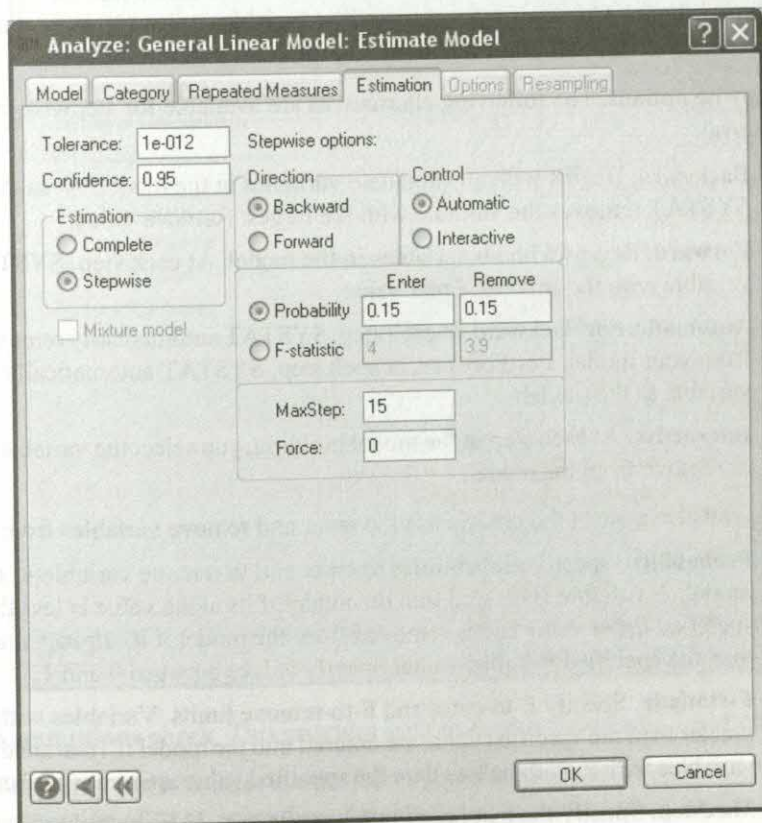
Levels. Number of repeated measures in the set. For example, suppose you have three dependent variables that represent measurements at different times, the number of levels is 3.

Metric. Metric that indicates the spacing between unevenly spaced measurements. For example, if measurements were taken at the third, fifth, and ninth weeks, the metric would be 3, 5, 9.

Estimation

The Estimation tab allows you to specify a tolerance and confidence level. You can select complete or stepwise estimation procedures and specify entry and removal criteria.

To specify estimation options, click the Estimation tab in the GLM: Estimate Model dialog box.



The following options can be specified:

Tolerance. Prevents the entry of a variable that is highly correlated with the independent variables already included in the model. Enter a value between 0 and 1.

Typical values are 0.01 or 0.001. The higher the value (closer to 1), the lower the correlation required to exclude a variable.

Confidence. Specify a confidence level for the confidence interval for the regression coefficients. The default level is 0.95.

Estimation. Controls the method used to enter and remove variables from the equation.

- **Complete.** All independent variables are entered in a single step.
- **Stepwise.** Variables are entered into or removed from the model, one at a time.

Mixture model. Constrains the independent variables to sum to a constant, when the Complete estimation option is chosen.

Stepwise options. The following alternatives are available for stepwise entry and removal:

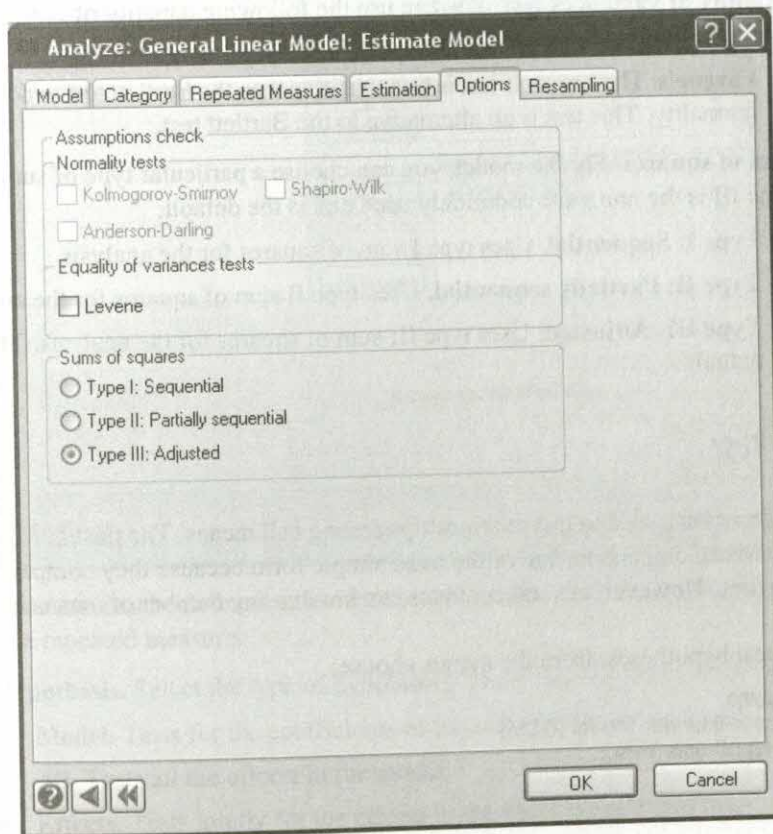
- **Backward.** Begins with all candidate variables in the model. At each step, SYSTAT removes the variable with the largest Remove value.
- **Forward.** Begins with no variables in the model. At each step, SYSTAT adds the variable with the smallest Enter value.
- **Automatic.** For Backward, at each step, SYSTAT automatically removes a variable from your model. For Forward, at each step, SYSTAT automatically adds a variable to the model.
- **Interactive.** At each step in the model building, you select the variable to enter into or remove from the model.

You can also control the criteria used to enter and remove variables from the model:

- **Probability.** Specify probabilities to enter and to remove variable(s) from the model. A variable is entered into the model if its alpha value is less than the specified Enter value and is removed from the model if its alpha value is greater than the specified Remove value. Specify values between 0 and 1.
- **F-statistic.** Specify F-to-enter and F-to-remove limits. Variables with F-statistic greater than the specified value are entered into the model if Tolerance permits and variables with F-statistic less than the specified value are removed from the model.
- **MaxStep.** Specify the maximum number of steps.
- **Force.** Forces the first n variables listed in your model to remain in the equation.

Options

To specify the options, click the Options tab in the General Linear Model: Estimate Model dialog box.



Assumptions check. This provides options to check the basic assumptions of GLM.

Normality tests. You can use the following normality tests to check the basic statistical assumption of GLM, normality of residuals.

- **Kolmogorov-Smirnov (Lillefors).** It is a nonparametric test used for large samples. It is applied to continuous distributions and gives greater importance to the observations in the centre than those at the tails.

- **Shapiro-Wilk.** The test provides the Shapiro-Wilk test statistic and p-value for the selected dependent variable. The smaller the p-value, the worse is the fit.
- **Anderson-Darling.** Anderson-Darling test is a standard goodness of fit test. It gives greater importance to the observations in the tails than those at the center.

Equality of variances test. You can use the following equality of variances test to check the homogeneity of variances across all levels of the factors:

- **Levene's.** The Levene's test is less sensitive than the Bartlett test to departures from normality. This test is an alternative to the Bartlett test.

Sum of squares. For the model, you can choose a particular type of sum of squares. Type III is the one most commonly used and is the default.

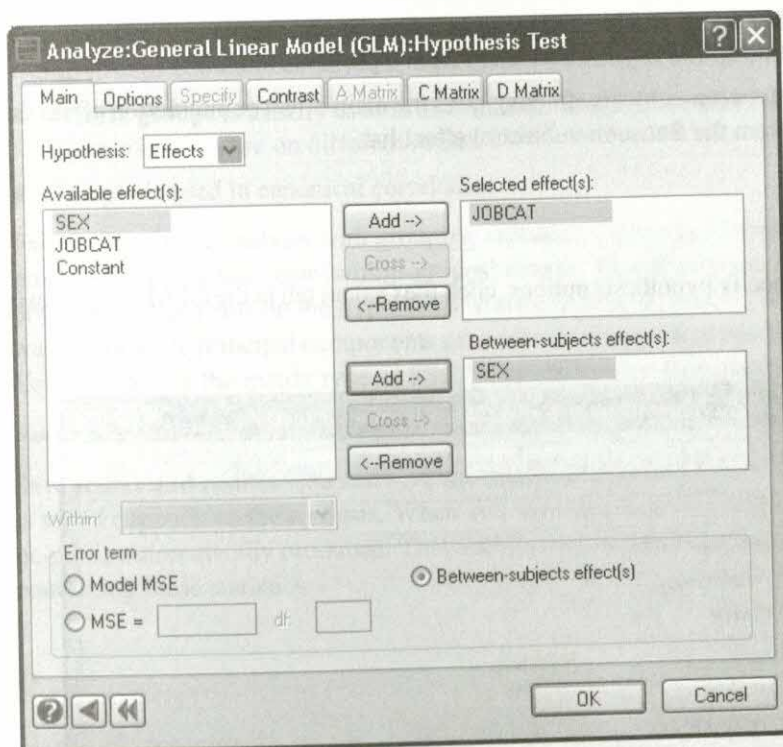
- **Type I: Sequential.** Uses type I sum of squares for the analysis.
- **Type II: Partially sequential.** Uses type II sum of squares for the analysis.
- **Type III: Adjusted.** Uses type III sum of squares for the analysis. This is the default.

Hypothesis Test

Contrasts are used to test relationships among cell means. The post hoc tests in GLM: Pairwise Comparisons are of the most simple form because they compare two means at a time. However, general contrasts can involve any number of means in the analysis.

To test hypotheses, from the menus choose:

Analyze
General Linear Model (GLM)
Hypothesis Test...



Contrasts can be defined across the categories of a grouping factor or across the levels of a repeated measure.

Hypothesis. Select the type of hypothesis. The following choices are available:

- **Model.** Tests for the coefficients of the model. This is the default.
- **All.** Tests all the effects in the model.
- **Effects.** Tests jointly for the effects in the Selected effect(s) list.
- **Specify.** Tests the hypotheses in the Specify tab.
- **A Matrix.** Tests the hypotheses corresponding to the A Matrix tab.

Within. Select the repeated measures factor across whose levels the contrast is defined.

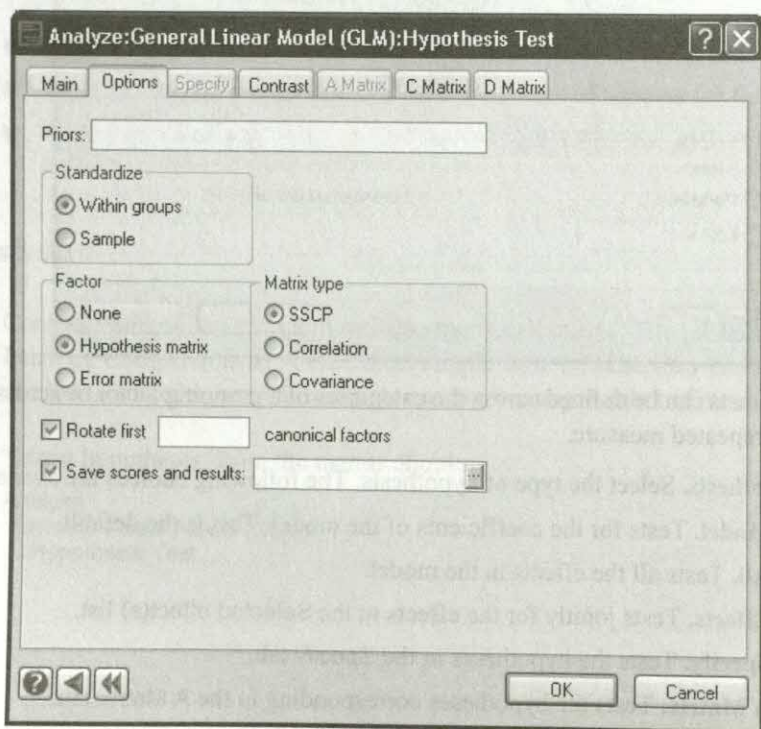
Error term. You can specify which error term to use for the hypothesis tests.

- **Model MSE.** Uses the mean square error from the general linear model that you ran.

- **MSE and df.** Uses the mean square error and degrees of freedom you specify. Use this option if you know them from a previous model.
- **Between-subjects effect(s).** Uses the main effect or interaction effect that you select from the Between-subject(s) effect list.

Options

To specify hypothesis options, click the Options tab in the GLM: Hypothesis Test dialog box.



Priors. Prior probabilities for discriminant analysis. Type a value for each group, separated by spaces. These probabilities should add to 1. For example, suppose you have three groups, priors might be 0.5, 0.3, and 0.2. The prior option is available when you select a single grouping variable as the effect to be tested.

Standardize. You can standardize canonical coefficients using the total sample or a within-groups covariance matrix.

- **Within groups** is usually used in discriminant analysis to make comparisons easier when measures are on different scales.
- **Sample** is used in canonical correlation.

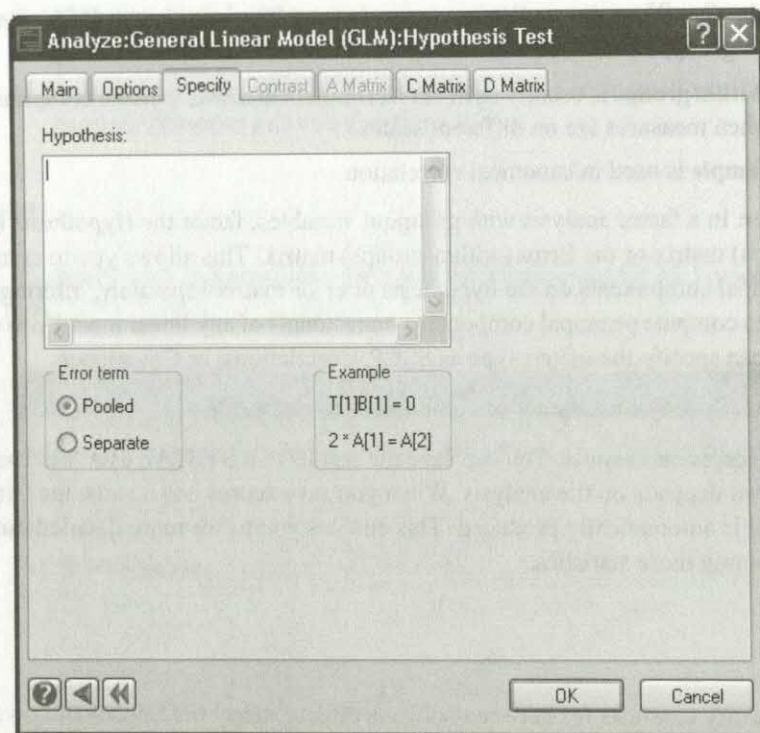
Factor. In a factor analysis with grouping variables, factor the Hypothesis (between-groups) matrix or the Error (within-groups) matrix. This allows you to compute principal components on the hypothesis or error matrix separately, offering a direct way to compute principal components on residuals of any linear model you wish to fit. You can specify the matrix type as SSCP, Correlations, or Covariance.

Rotate. Specify the number of components to rotate.

Save scores and results. You can save the results to a SYSTAT data file. Exactly what is saved depends on the analysis. When you save scores and results, the extended output is automatically produced. This enables you to see more detailed output when computing these statistics.

Specify

To specify contrasts for between-subjects effects, select the Specify option of Hypothesis in the GLM: Hypothesis Test dialog box. The Specify tab gets enabled.



You can use GLM's cell means "language" to define contrasts across the levels of a grouping variable in a multivariate model. For example, for a two-way factorial ANOVA design with *DISEASE* (four categories) and *DRUG* (three categories), you could contrast the marginal mean for the first level of drug against the third level by specifying:

$$DRUG[1] = DRUG[3]$$

Note that square brackets enclose the value of the category. For string variables, their values are assumed to be in the upper case unless they are enclosed in quotes. For example, *GENDER*\$(male) is read as *GENDER*\$(MALE), whereas *GENDER*\$['male'] will prompt SYSTAT to look for the exact string 'male'. For the simple contrast of the first and third levels of *DRUG* for the second disease only, specify:

$$DRUG[1] \text{ DISEASE}[2] = DRUG[3] \text{ DISEASE}[2]$$

The syntax also allows statements like:

$$-3*DRUG[1] - 1*DRUG[2] + 1*DRUG[3] + 3*DRUG[4]$$

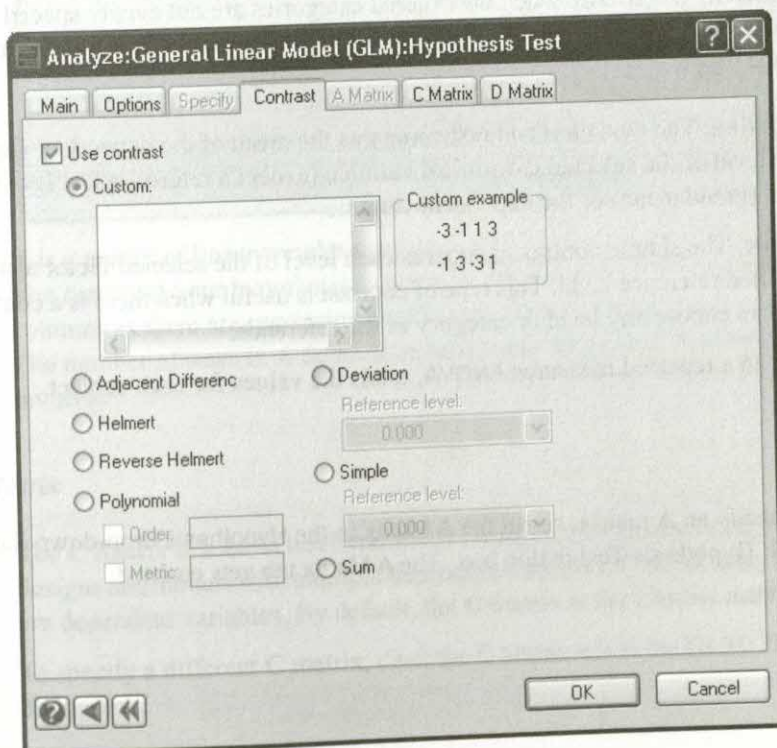
where the right-hand side is considered zero unless you specify a value for it or specify it through a **D** matrix. For a univariate model, you can also choose one of the following:

Pooled. Uses the error term from the current model.

Separate. Generates a separate variances error term.

Contrast

Contrast generates a contrast for a grouping factor or a repeated measures factor. To specify contrasts, select an effect under the Effect option of Hypothesis or a repeated measures factor in the Within drop-down list. The Contrast tab gets enabled.



SYSTAT offers eight types of contrasts:

Custom. Enter your own custom coefficients. For example, suppose your factor has four ordered categories (or levels), you can specify your own coefficients, such as $-3 -1 1 3$, by typing these values in the Custom text box.

Adjacent difference. Compares each level of the factor with its adjacent level.

Helmert. Compares the mean of each level of the selected factor with the mean of the succeeding levels.

Reverse Helmert. Compares the mean of each level of the selected factor with the mean of the previous levels.

Polynomial. Generates orthogonal polynomial contrasts (to test linear, quadratic, or cubic trends across ordered categories or levels).

- **Order.** Enter 1 for linear, 2 for quadratic, etc.
- **Metric.** Use Metric when the ordered categories are not evenly spaced. For example, when repeated measures are collected at weeks 2, 4, and 8, enter 2,4,8 as the metric.

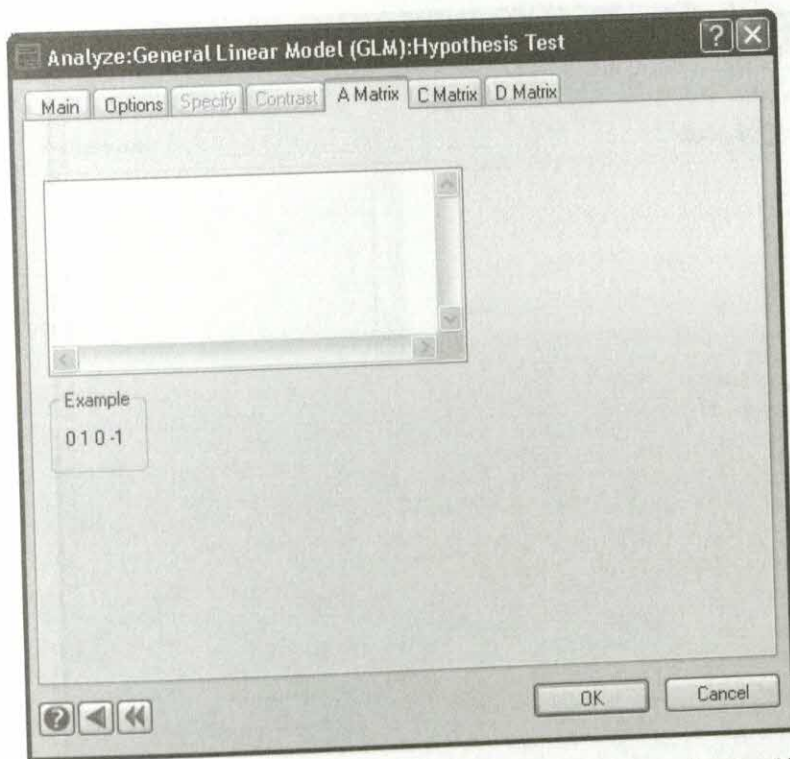
Deviation. The deviation contrast compares the mean of the dependent variable for each level of the selected categorical variable (except a reference level) to the overall mean (grand mean) of the dependent variable.

Simple. The simple contrast compares each level of the selected factor against the specified reference level. This type of contrast is useful when there is a control group. You can choose any level or category as the reference.

Sum. In a repeated measures ANOVA, totals the values for each subject.

A Matrix

To specify an A matrix, select the A Matrix in the Hypothesis drop-down list of the GLM: Hypothesis Test dialog box. The A Matrix tab gets enabled.

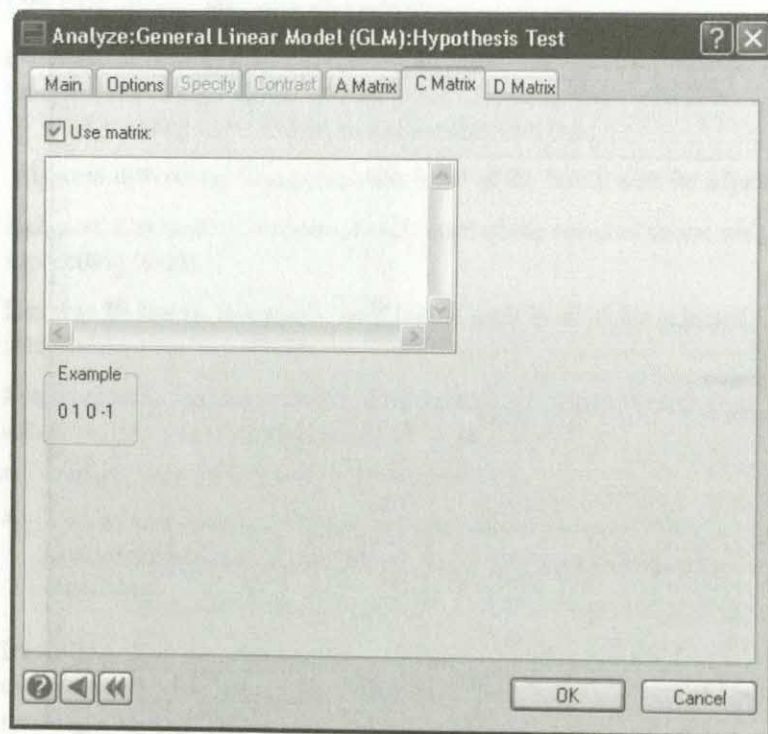


A is a matrix of linear weights contrasting the coefficient estimates (the rows of **B**). You can write your hypothesis in terms of the **A** matrix. The **A** matrix has as many columns as there are regression coefficients (including the constant) in your model. The number of rows in **A** determines how many degrees of freedom your hypothesis involves.

C Matrix

The **C** matrix is used to test hypotheses for repeated measures analysis of variance designs and models with multiple dependent variables. **C** has as many columns as there are dependent variables. By default, the **C** matrix is the identity matrix.

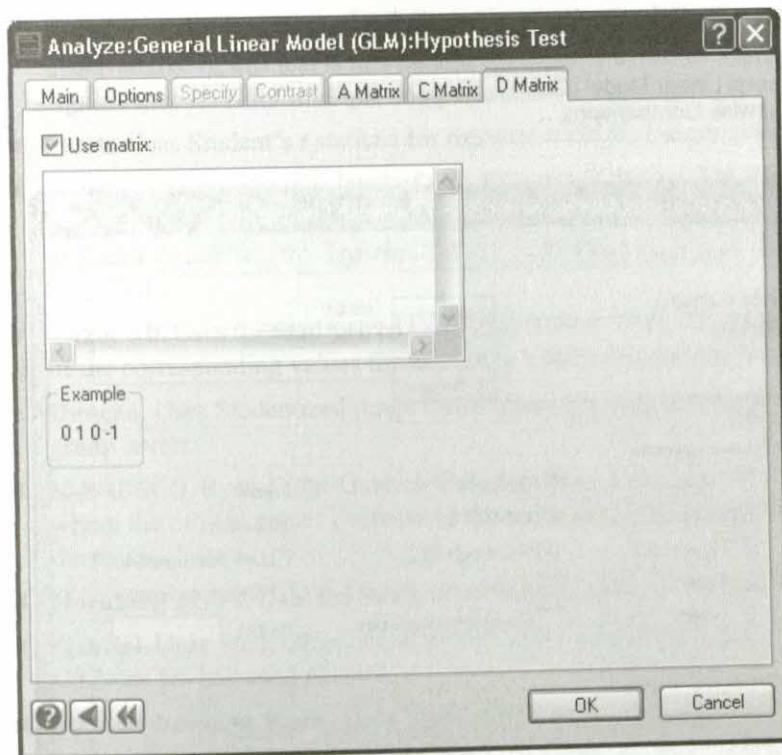
To specify a different **C** matrix, click the **C Matrix** tab in the GLM: Hypothesis Test dialog box.



D Matrix

D is a null hypothesis matrix (by default null matrix). The **D** matrix, if you use it, must have the same number of rows as **A**. For univariate multiple regression, **D** has only one column. For multivariate models (multiple dependent variables), the **D** matrix has one column for each dependent variable.

To specify a different **D** matrix, click the **D** matrix tab in the GLM: Hypothesis Test dialog box.



Toggling between the command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through a dialog box, then post estimation analysis can be performed through commands and vice-versa.

Pairwise Comparisons

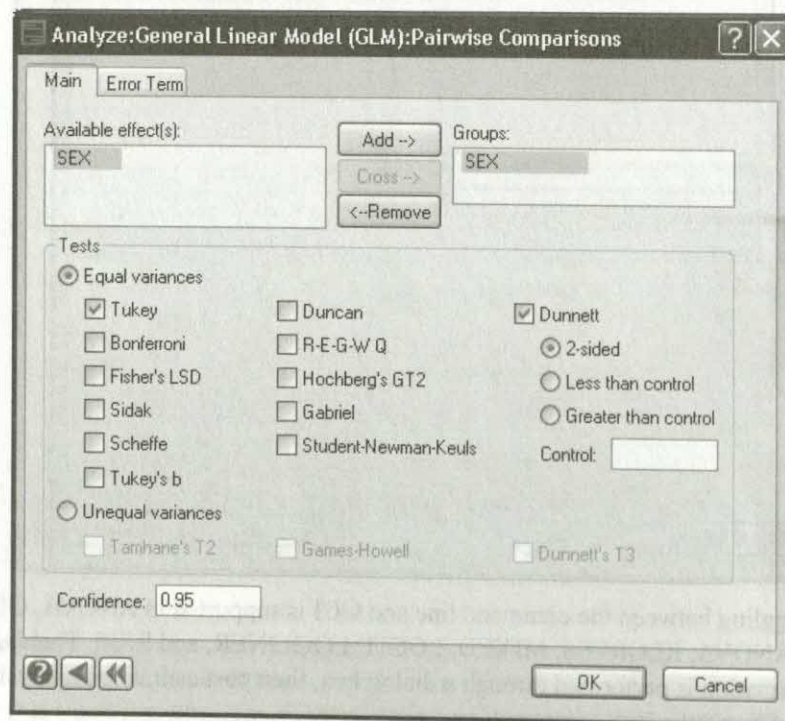
After fitting the model, one can find the treatment pairs which are significantly different, or form several homogeneous sets of treatments with their respective *p-values* by using several multiple comparison tests (mct) offered by SYSTAT under equal or unequal variance assumptions.

To open the Pairwise Comparisons dialog box, from the menus choose:

Analyze

General Linear Model (GLM)

Pairwise Comparisons...



Groups. Select the variable that defines the groups.

Tests. There are several post hoc tests to compare the means of the dependent variable for the selected grouping variable.

Equal variances. Tests in this group assume equality of variances across all levels of the grouping variable.

- **Tukey.** Uses the Studentized range distribution to make all pairwise comparisons. This is the default.
- **Bonferroni.** Uses Student's t statistic. It sets the family-wise error rate as $(1 - \text{Confidence}) / (\text{Total number of comparisons})$.

- **Fisher's LSD.** Equivalent to multiple t tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.
- **Sidak.** Uses Student's t statistic for pairwise multiple comparisons.
- **Scheffé.** The significance level of Scheffé's test is designed to allow all possible linear combinations of group means to be tested, not just pairwise comparisons available in this feature. The result is that Scheffé's test is more conservative than other tests.
- **Tukey's b.** Uses the Studentized range distribution. The critical value is the average of the corresponding values for the Tukey's HSD test and the S-N-K test.
- **Duncan.** Uses Studentized range distribution. It yields homogeneous subsets of group levels.
- **R-E-G-W Q.** Ryan-Einot-Gabriel-Welsch Q test is a modification of the S-N-K test where the critical values decrease as the range in the set being considered decreases.
- **Hochberg's GT2.** Uses the Studentized maximum modulus distribution.
- **Gabriel.** Uses Studentized maximum modulus distribution. It is equivalent to the GT2 test for balanced ANOVA.
- **Student-Newman-Keuls.** Uses Studentized range distribution. It yields homogenous subsets of group levels.
- **Dunnett.** The Dunnett test is available only with one-way designs. Dunnett compares a set of treatments against a single control mean that you specify. You can choose from the following three alternative hypotheses: (a) 2-sided (not equal to), (b) less than, or (c) greater than control level. 2-sided is the default.

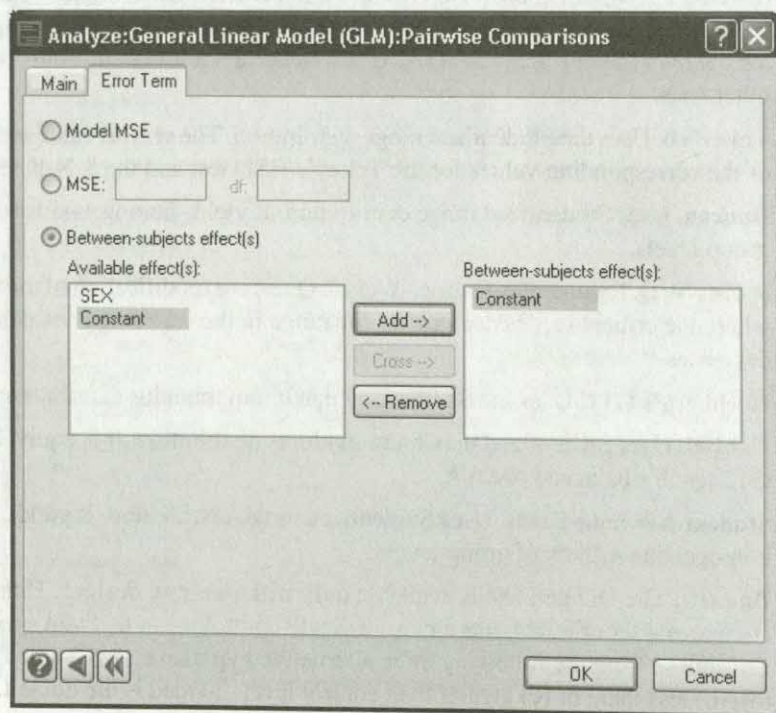
Unequal variances. The following tests do not require the homogeneity of variance assumption. These tests use the Welsch procedure for determining the denominator degrees of freedom.

- **Tamhane's T2.** Uses the Student's t distribution. Uses the Sidak inequality to find the alpha level.
- **Games-Howell.** Uses the Studentized range distribution.
- **Dunnett's T3.** Uses the Studentized maximum modulus distribution.

Confidence. Specify confidence level for pairwise comparisons tests. The default value is 0.95.

Error term

To specify the error term, click the Error Term tab in the GLM: Pairwise Comparisons dialog box.



Error term. You can choose one of the following:

- **Model MSE.** Uses the mean square error (MSE) from the general linear model that you ran.
- **MSE and df.** Uses the mean square error term and degrees of freedom that you specify. Use this option if you know them from a previous model.
- **Between-subjects effect(s).** Select this option to use the main effect error term or the interaction error term in all the tests.

Toggling between the command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if

estimation is performed through a dialog box then post estimation analysis can be performed through commands and vice-versa.

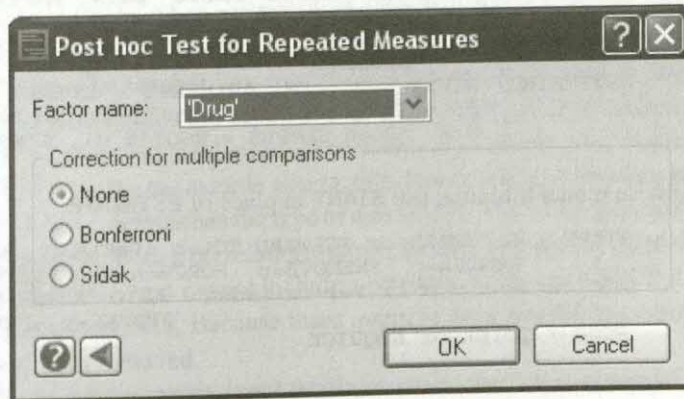
Post hoc Tests for Repeated Measures

After performing analysis of variance, we just have an *F-ratio*, which tells us that means are not equal--we still do not know exactly which means are significantly different from which other ones. Post hoc tests can only be used when the "omnibus" ANOVA found a significant effect. If the *F-value* for a factor turns out non-significant, you cannot go further with the analysis. This protects the post hoc test from being used too liberally.

The main problem that designers of post hoc tests try to deal with is alpha inflation. This refers to the fact that the more tests you conduct at $\alpha=0.05$, the more likely you are to claim you have significant result when you shouldn't have. The overall chance of a type I error rate in a particular experiment is referred to as the "experiment-wise error rate" (or family-wise error rate).

To perform the Post hoc Test for Repeated Measures, from the menus choose:

Analyze
General Linear Model (GLM)
Post hoc Test for Repeated Measures...



Factor name. Select a factor name from the drop-down list of factors defined for the model.

Correction for multiple comparisons. The following options are available:

- **Bonferroni.** To keep the experiment-wise error rate to a specified level ($\alpha=0.05$) a simple way is to divide the acceptable α level by the number of comparisons we intend to make. That is, for any one comparison to be considered significant, the obtained p -value would have to be less than $\alpha/(\text{num of comparisons})$. Select this option to perform a Bonferroni correction.
- **Sidak.** The experiment-wise error explained above is kept in control by the use of the formula: $\text{sidak_alpha} = 1 - (1 - \alpha)^{1/c}$, where c is the number of paired comparisons.

Toggling between the command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through a dialog box then post estimation analysis can be performed through commands and vice-versa.

Using Commands

Select the data with USE FILENAME and continue with:

```
GLM
  MODEL varlist1=CONSTANT + varlist2 + var1*var2 +, var3(var4)/
    REPEAT=m,n,... REPEAT=m(x1,x2,...), n(y1,y2,...)
    NAMES='name1','name2',... , MEANS, WEIGHT N=n
  CATEGORY grpvarlist / MISS EFFECT or DUMMY
  PLENGTH SHORT or MEDIUM or LONG
  SAVE filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
    'comment'
  WORK filename / COEF MODEL RESID DATA PARTIAL ADJUSTED
    'comment'
  ESTIMATE / NTEST = KS, SW, AD HTEST = LEVENE
    SS = TYPE1 or TYPE2 or TYPE3 QUICK or NOQUICK
    MIX TOL=n SAMPLE = BOOT(m,n), SIMPLE(m,n), JACK
```

For stepwise model building, use START in place of ESTIMATE:

```
START / BACKWARD or FORWARD TOL=n ENTER=p REMOVE=p,
    FENTER=n FREMOVE=n FORCE=n MAXSTEP=n
STEP no argument or var or index / AUTO ENTER=p, REMOVE=p
    FENTER=n FREMOVE=n
STOP / QUICK or NOQUICK
```


To perform hypothesis tests:

```

HYPOTHESIS
  EFFECT varlist, var1&var2,...
  WITHIN 'name'
  CONTRAST [matrix] / DIFFERENCE or SUM or DEVIATION [c] or
                        SIMPLE[c] or HELMERT or RHELMERT or
                        POLYNOMIAL ORDER=n METRIC=m,n,...
  SPECIFY hypothesis lang / POOLED or SEPARATE
  AMATRIX [matrix]
  CMATRIX [matrix]
  DMATRIX [matrix]
  ALL
  POST grpvar/ LSD BONF=n TUKEY SCHEFFE SIDAK SNK BTUKEY DUNCAN
                GT2 GABR QREG GH T2 T3 POOLED SEPARATE
                DUNNETT = LT or GT or TWO CONTROL = 'levelname'
  PAIRWISE 'factorname' / BONF or SIDAK
  ROTATE n
  TYPE CORR or COVAR or SSCP
  STAND TOTAL or WITHIN
  FACTOR HYPOTHESIS or ERROR
  ERROR value(df) or var or var1*var2 or var1 & var2 or matrix
  PRIORS m n p ...
  TEST/CONFI = n

```

Usage Considerations

Types of data. Normally, you analyze raw cases-by-variables data with GLM. You can, however, use a symmetric matrix data file (for example, a covariance matrix saved in a file from Correlations) as input. Suppose you use a matrix as input, you must specify a value for Cases when estimating the model (under the Model tab in the GLM: Estimate Model dialog box) to specify the sample size of the data file that generated the matrix. The number you specify must be an integer greater than two.

Be sure to include the dependent as well as independent variables in your matrix. SYSTAT picks out the dependent variable you name in your model.

SYSTAT uses the sample size to calculate degrees of freedom in hypothesis tests. SYSTAT also determines the type of matrix (SSCP, Covariance, and so on) and adjusts appropriately. With a correlation matrix, the raw and standardized coefficients are the same; therefore, you cannot include a constant when using SSCP, Covariance, or Correlation matrices. Because these matrices are centered, the constant term has already been removed.

The triangular matrix input facility is useful for “meta-analysis” of published data and missing value computations; however, you should heed the following warnings: First, suppose you input correlation matrices from textbooks or articles, you may not get the same regression coefficients as those printed in the source. Because of round-

off error, printed and raw data can lead to different results. Second, suppose you use pairwise deletion with Correlations, the degrees of freedom for hypotheses will not be appropriate. You may not even be able to estimate the regression coefficients because of singularities.

In general, correlation matrices containing missing data produce coefficient estimates and hypothesis tests that are optimistic. You can correct for this by specifying a sample size smaller than the number of actual observations (preferably set it equal to the smallest number of cases used for any pair of variables), but this is a guess that you can refine only by doing Monte Carlo simulations. There is no simple solution. Beware, especially, of multivariate regressions (MANOVA and others) with missing data on the dependent variables. You can usually compute coefficients, but hypothesis testing produces results that are suspect.

Print options. GLM produces extended output if you set the output length to LONG or if you select Save scores and results in the GLM Hypothesis Test dialog box.

For model estimation, the extended output adds the following: total sum of product matrix, residual (or pooled within groups) sum of product matrix, residual (or pooled within groups) covariance matrix, and the residual (or pooled within groups) correlation matrix.

For hypothesis testing, the extended output adds **A**, **C**, and **D** matrices, the matrix of contrasts, and the inverse of the cross products of contrasts, hypothesis and error sum of product matrices, tests of residual roots, canonical correlations, coefficients, and loadings.

Quick Graphs. If no variables are categorical, GLM produces Quick Graphs of residuals versus predicted values. For categorical predictors, GLM produces graphs of the least-squares means for the levels of the categorical variable(s).

Saving files. Several sets of output can be saved to a file. The actual contents of the saved file depend on the analysis. Files may include estimated regression coefficients, model variables, residuals, predicted values, diagnostic statistics, canonical variable scores, and posterior probabilities (among other statistics).

BY groups. Each level of a BY variable yields a separate analysis. However, for Hypothesis Testing, BY groups does not work. You have to resort to Data--> Select Cases commands.

Case frequencies. GLM uses the FREQUENCY variable, if present, to duplicate cases.

Case weights. GLM uses the values of any WEIGHT variable to weight each case.

Examples

Example 1 One-Way ANOVA

The following data, *KENTON*, are from Kutner et al. (2004). The data comprise unit sales of a cereal product under different types of package designs. Ten stores were selected as experimental units. Each store was randomly assigned to sell one of the package designs (each design was sold at two or three stores).

PACKAGE SALES

1	12
1	18
2	14
2	12
2	13
3	19
3	17
3	21
4	24
4	30

Numbers are used to code the four types of package designs; alternatively, you could have used words. Kutner et al. (2004) report that cartoons are part of designs 1 and 3 but not designs 2 and 4; designs 1 and 2 have three colors; and designs 3 and 4 have five colors. Thus, string codes for *PACKAGE\$* might have been 'Cart 3', 'NoCart 3', 'Cart 5', and 'NoCart 5'. Notice that the data does not need to be ordered by *PACKAGE* as shown here.

The input is:

```
GLM
USE KENTON
CATEGORY PACKAGE
MODEL SALES=CONSTANT + PACKAGE
GRAPH NONE
ESTIMATE
```

The output is:

Effects coding used for categorical variables in model.
Categorical values encountered during processing are

Variables	Levels
PACKAGE (4 levels)	1.000 2.000 3.000 4.000
Dependent Variable	SALES
N	10
Multiple R	0.921
Squared Multiple R	0.849

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
PACKAGE	258.000	3	86.000	11.217	0.007
Error	46.000	6	7.667		

This is the standard analysis of variance table. The *F-ratio* (11.217) appears significant, so you could conclude that the package designs differ significantly in their effects on sales, provided the assumptions are valid.

Pairwise Multiple Comparisons

SYSTAT offers fifteen methods for comparing pairs of means: Bonferroni, Tukey-Kramer HSD, Sidak, Duncan, Scheffé, Fisher's LSD, Tukey's B, R-E-G-W Q, Hochberg's GT2, Tamhane's T2, Games-Howell, Dunnett T3, Gabriel, Student-Newman-Keuls, and Dunnett's test.

The Dunnett test is available only with one-way designs. Dunnett requires the value of a control group against which comparisons are made. By default, two-sided tests are computed. One-sided Dunnett tests are also available. Incidentally, for Dunnett's tests on experimental data, you should use the one-sided option unless you cannot predict from theory whether your experimental groups will have higher or lower means than the control.

Comparisons for the pairwise methods are made across all pairs of least-squares group means for the design term that is specified. For a multiway design, marginal cell means are computed for the effects specified before the comparisons are made.

To determine significant differences, simply look for pairs with probabilities below your critical value (for example, 0.05 or 0.01). All multiple comparison methods handle unbalanced designs correctly.

After you estimate your ANOVA model, it is easy to do post hoc tests. To do a Tukey-Kramer HSD test, first estimate the model, then specify these commands.

The input is:

```
HYPOTHESIS
POST PACKAGE / TUKEY
TEST
```

The output is:

Post Hoc Test of SALES
Using least squares means.
Using model MSE of 7.667 with 6 df.

Tukey's Honestly-Significant-Difference Test

PACKAGE (i)	PACKAGE (j)	Difference	p-value	95.0% Confidence Interval	
				Lower	Upper
1	2	2.000	0.856	-6.750	10.750
1	3	-4.000	0.452	-12.750	4.750
1	4	-12.000	0.019	-21.585	-2.415
2	3	-6.000	0.130	-13.826	1.826
2	4	-14.000	0.006	-22.750	-5.250
3	4	-8.000	0.071	-16.750	0.750

Results show that sales for the fourth package design (five colors and no cartoons) are significantly larger than those for packages 1 and 2. None of the other pairs differ significantly.

Contrasts

This example uses two contrasts:

- We compare the first and third packages using coefficients of (1, 0, -1, 0).
- We compare the average performance of the first three packages with the last, using coefficients of (1, 1, 1, -3).

The input is:

```
HYPOTHESIS
EFFECT PACKAGE
CONTRAST [1 0 -1 0]
TEST
```

```
HYPOTHESIS
EFFECT PACKAGE
CONTRAST [1 1 1 -3]
TEST
```

For each hypothesis, we specify one contrast, so the test has one degree of freedom; therefore, the contrast matrix has one row of numbers. These numbers are the same

ones you see in ANOVA textbooks, although ANOVA offers one advantage—you do not have to standardize them so that their sum of squares is 1.

The output is:

Test for effect called: PACKAGE

A Matrix

	1	2	3	4
	0.000	1.000	0.000	-1.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	19.200	1	19.200	2.504	0.165
Error	46.000	6	7.667		

Test for effect called: PACKAGE

A Matrix

	1	2	3	4
	0.000	4.000	4.000	4.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	204.000	1	204.000	26.609	0.002
Error	46.000	6	7.667		

For the first contrast, the *F-ratio* (2.504) is not significant, so you cannot conclude that the impact of the first and third package designs on sales is significantly different. Incidentally, the **A** matrix contains the contrast. The first column (0) corresponds to the constant in the model, and the remaining three columns (1 0 -1) correspond to the dummy variables for *PACKAGE*.

The last package design is significantly different from the other three taken as a group. Notice that the **A** matrix looks much different this time. Because the effects sum to 0, the last effect is minus the sum of the other three; that is, letting α_i denote the effect for level *i* of package,

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 0$$

so

$$\alpha_4 = -(\alpha_1 + \alpha_2 + \alpha_3)$$

and the contrast is

$$\alpha_1 + \alpha_2 + \alpha_3 - 3\alpha_4$$

which is

$$\alpha_1 + \alpha_2 + \alpha_3 - 3(-\alpha_1 - \alpha_2 - \alpha_3)$$

which simplifies to

$$4\alpha_1 + 4\alpha_2 + 4\alpha_3$$

Remember, SYSTAT does all this work automatically.

Orthogonal Polynomials

Constructing orthogonal polynomials for between-group factors is useful when the levels of a factor are ordered. To construct orthogonal polynomials for your between-groups factors, the input is:

```
HYPOTHESIS
EFFECT PACKAGE
CONTRAST / POLYNOMIAL ORDER=2
TEST
```

The output is:

Test for effect called: PACKAGE

A Matrix

1	2	3	4
0.000	0.000	-1.000	-1.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	60.000	1	60.000	7.826	0.031
Error	46.000	6	7.667		

Make sure that the levels of the factor—after they are sorted by the procedure numerically or alphabetically—are ordered meaningfully on a latent dimension. Suppose you need a specific order, use LABEL or ORDER; otherwise, the results will not make sense. In the example, the significant quadratic effect is the result of the fourth package having a much larger sales volume than the other three.

Effect and Dummy Coding

The effects in a least-squares analysis of variance are associated with a set of dummy variables that SYSTAT generates automatically. Ordinarily, you do not have to concern yourself with these dummy variables; suppose you want to see them, you can save them in to a SYSTAT file.

The input is:

```
GLM
  USE KENTON
  CATEGORY PACKAGE
  MODEL SALES=CONSTANT + PACKAGE
  GRAPH NONE
  SAVE MYCODES / MODEL
  ESTIMATE
  USE MYCODES
  FORMAT 12,0
  LIST SALES x( 1.. 3)
```

The listing of the dummy variables follows.

The output is:

Case	SALES	X(1)	X(2)	X(3)
1	12	1	0	0
2	18	1	0	0
3	14	0	1	0
4	12	0	1	0
5	13	0	1	0
6	19	0	0	1
7	17	0	0	1
8	21	0	0	1
9	24	-1	-1	-1
10	30	-1	-1	-1

The variables $X(1)$, $X(2)$, and $X(3)$ are the effects-coding dummy variables generated by the procedure. All cases in the first cell are associated with dummy values 1 0 0; those in the second cell with 0 1 0; the third, 0 0 1; and the fourth, -1 -1 -1. Other least-squares programs use different methods to code dummy variables. The coding used by SYSTAT is the one most widely used and guarantees that the effects sum to 0.

If you had used dummy coding, these dummy variables would be saved:

SALES	X(1)	X(2)	X(3)
12	1	0	0
18	1	0	0
14	0	1	0
12	0	1	0
13	0	1	0
19	0	0	0
19	0	0	1
17	0	0	1
21	0	0	1
24	0	0	0
30	0	0	0

This coding yields parameter estimates that are the differences between the mean for each group and the mean of the last group.

Example 2

Analysis of Covariance (ANCOVA)

Winer, Brown and Michels (1991) use the *COVAR* data file for an analysis of covariance in which *X* is the covariate and *TREAT* is the treatment. Cases do not need to be ordered by the grouping variable *TREAT*.

To define an ANCOVA model in GLM, we have to select factors (*TREAT*) and covariates (*X*) as independent variables and define only factors as categorical variable(s). SYSTAT automatically assumes non-categorical variables as covariates.

The input is:

```
GLM
  USE COVAR
  CATEGORY TREAT
  MODEL Y = CONSTANT + TREAT + X + TREAT*X
  ESTIMATE
```


The output is:

```
Dependent Variable : Y
N : 21
Multiple R : 0.921
Squared Multiple R : 0.849
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
TREAT	6.693	2	3.346	5.210	0.019
X	15.672	1	15.672	24.399	0.000
TREAT*X	0.667	2	0.334	0.519	0.605
Error	9.635	15	0.642		

The probability value for the treatment by covariate interaction is 0.605, so the assumption of homogeneity of slopes is justifiable.

Testing Interaction Contrasts

One might be interested in testing the interaction effect for different levels of treatment. SYSTAT does not support interaction contrasts directly through the CONTRAST command but you can test this by using the AMATRIX command.

The input is:

```
NOTE "INTERACTION CONTRAST [1 0 -1]"
HYPOTHESIS
AMATRIX [0 0 0 0 2 1]
TEST
```

```
NOTE "INTERACTION CONTRAST [-2 1 1]"
HYPOTHESIS
AMATRIX [0 0 0 0 -3 0]
TEST
```

The output is:

Interaction Contrast [1 0 -1]

A Matrix

1	2	3	4	5
0.000	0.000	0.000	0.000	2.000

A Matrix

6
1.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	0.558	1	0.558	0.868	0.366
Error	9.635	15	0.642		

Interaction Contrast [-2 1 1]

A Matrix

1	2	3	4	5
0.000	0.000	0.000	0.000	-3.000

A Matrix

6
0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	0.665	1	0.665	1.036	0.325
Error	9.635	15	0.642		

Notice that the interaction contrast matrix and the A Matrix are different. Refer Example 1 in "Contrasts" on page 205

Example 3

Randomized Block Designs

A randomized block design is like a factorial design without an interaction term. The following example is from Kutner et al. (2004). Five blocks of judges were given the task of analyzing three treatments. Judges are stratified within blocks, so the interaction of blocks and treatments cannot be analyzed. These data are in the file *BLOCK*.

The input is:

```
GLM
  USE BLOCK
  CATEGORY BLOCK, TREAT
  MODEL JUDGMENT = CONSTANT + BLOCK + TREAT
  ESTIMATE
```

You must use GLM instead of ANOVA because you do not want the *BLOCK*TREAT* interaction in the model.

The output is:

Dependent Variable	JUDGMENT
N	15
Multiple R	0.970
Squared Multiple R	0.940

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
BLOCK	171.333	4	42.833	14.358	0.001
TREAT	202.800	2	101.400	33.989	0.000
Error	23.867	8	2.983		

Example 4

Incomplete Block Designs

Randomized blocks can be used in factorial designs. Here is an example from John (1971). The data (in the file *JOHN*) involve an experiment with three treatment factors (*A*, *B*, and *C*) plus a blocking variable with eight levels. Notice that data were collected on 32 of the possible 64 experimental situations.

BLOCK	A	B	C	Y	BLOCK	A	B	C	Y
1	1	1	1	101	5	1	1	1	87
1	2	1	2	373	5	2	1	2	324
1	1	2	2	398	5	1	2	1	279
1	2	2	1	291	5	2	2	2	471
2	1	1	2	312	6	1	1	2	323
2	2	1	1	106	6	2	1	1	128
2	1	2	1	265	6	1	2	2	423
2	2	2	2	450	6	2	2	1	334
3	1	1	1	106	7	1	1	1	131
3	2	2	1	306	7	2	1	1	103
3	1	1	2	324	7	1	2	2	445
3	2	2	2	449	7	2	2	2	437
4	1	2	1	272	8	1	1	2	324
4	2	1	1	89	8	2	1	2	361
4	1	2	2	407	8	1	2	1	302
4	2	1	2	338	8	2	2	1	272

The input is:

```
GLM
USE JOHN
CATEGORY BLOCK, A, B, C
MODEL Y = CONSTANT + BLOCK + A # B # C
ESTIMATE
```

The output is:

```
Dependent Variable : Y
N : 32
Multiple R : 0.994
Squared Multiple R : 0.988
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
BLOCK	2638.469	7	376.924	1.182	0.364
A	3465.281	1	3465.281	10.862	0.004
B	161170.031	1	161170.031	505.209	0.000
C	278817.781	1	278817.781	873.992	0.000
A*B	28.167	1	28.167	0.088	0.770
A*C	1802.667	1	1802.667	5.651	0.029
B*C	11528.167	1	11528.167	36.137	0.000
A*B*C	45.375	1	45.375	0.142	0.711
Error	5423.281	17	319.017		

Example 5 Fractional Factorial Designs

Sometimes a factorial design involves so many combinations of treatments that certain cells must be left empty to save experimental resources. At other times, a complete randomized factorial study is designed, but loss of subjects leaves one or more cells completely missing. These models are similar to incomplete block designs because not all effects in the full model can be estimated. Usually, certain interactions must be left out of the model.

The following example uses some experimental data that contain values in only 8 out of 16 possible cells. Each cell contains two cases. The pattern of non-missing cells

makes it possible to estimate only the main effects plus three two-way interactions. The data are in the file *FRACTION*.

A	B	C	D	Y
1	1	1	1	7
1	1	1	1	3
2	2	1	1	1
2	2	1	1	2
2	1	2	1	12
2	1	2	1	13
1	2	2	1	14
1	2	2	1	15
2	1	1	2	8
2	1	1	2	6
1	2	1	2	12
1	2	1	2	10
1	1	2	2	6
1	1	2	2	4
2	2	2	2	6
2	2	2	2	7

The input is:

```
GLM
  USE FRACTION
  CATEGORY A, B, C, D
  MODEL Y = CONSTANT + A + B + C + D + A*B + A*C + B*C
  ESTIMATE
```

We must use GLM instead of ANOVA to omit the higher-way interactions that ANOVA automatically generates.

The output is:

```
Dependent Variable : Y
N : 16
Multiple R : 0.972
Squared Multiple R : 0.944
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
A	16.000	1	16.000	8.000	0.022
B	4.000	1	4.000	2.000	0.195
C	49.000	1	49.000	24.500	0.001
D	4.000	1	4.000	2.000	0.195
A*B	182.250	1	182.250	91.125	0.000
A*C	12.250	1	12.250	6.125	0.038
B*C	2.250	1	2.250	1.125	0.320
Error	16.000	8	2.000		

When missing cells turn up by chance rather than by design, you may not know which interactions to eliminate. When you attempt to fit the full model, SYSTAT informs you that the design is singular. In that case, you may need to try several models before finding an estimable one. It is usually best to begin by leaving out the highest-order interaction (A*B*C*D in this example). Continue with subset models until you get an ANOVA table.

Looking for an estimable model is not the same as analyzing the data with stepwise regression because you are not looking at *p-values*. After you find an estimable model, stop and settle with the statistics printed in the ANOVA table.

Example 6

Nested Designs

Nested designs resemble factorial designs with certain cells missing (incomplete factorials). This is because one factor is nested under another, so not all combinations of the two factors are observed. For example, in an educational study, classrooms are usually nested under schools because it is impossible to have the same classroom existing at two different schools (except as antimatter). The following example (in which teachers are nested within schools) is from Kutner et al. (2004). The data (learning scores) look like this:

	TEACHER1	TEACHER2
SCHOOL1	25	14
	29	11
SCHOOL2	11	22
	6	18
SCHOOL3	17	5
	20	2

In the study, there are actually six teachers, not just two; thus, the design really looks like this:

	TEACHER1	TEACHER2	TEACHER3	TEACHER4	TEACHER5	TEACHER6
SCHOOL1	25 29	14 11				
SCHOOL2			11 6	22 18		
SCHOOL3					17 20	5 2

The data are set up in the file *SCHOOLS*

TEACHER	SCHOOL	LEARNING
1	1	25
1	1	29
2	1	14
2	1	11
3	2	11
3	2	6
4	2	22
4	2	18
5	3	17
5	3	20
6	3	5
6	3	2

The input is:

GLM

USE SCHOOLS

CATEGORY TEACHER, SCHOOL

MODEL LEARNING = CONSTANT + SCHOOL + TEACHER(SCHOOL)

ESTIMATE

The output is:

```

Dependent Variable : LEARNING
N                  : 12
Multiple R         : 0.972
Squared Multiple R : 0.945

```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
SCHOOL	156.500	2	78.250	11.179	0.009
TEACHER (SCHOOL)	567.500	3	189.167	27.024	0.001
Error	42.000	6	7.000		

Your data can use any codes for *TEACHER*, including a separate code for every teacher in the study, as long as each teacher within a given school has a different code. GLM will use the nesting specified in the MODEL statement to determine the pattern of nesting. You can, for example, allow teachers in different schools to share codes.

This example is a balanced nested design. Unbalanced designs (unequal number of cases per cell) are handled automatically in SYSTAT because the estimation method is least-squares.

Example 7

Split Plot Designs

The split plot design is closely related to the nested design. In the split plot, however, plots are often considered a random factor; therefore, you have to construct different error terms to test different effects. The following example involves two treatments: *A* (between plots) and *B* (within plots). The numbers in the cells are the *YIELD* of the crop within plots.

	A1		A2	
	PLOT1	PLOT2	PLOT3	PLOT4
B1	0	3	4	5
B2	0	1	2	4
B3	5	5	7	6
B4	3	4	8	6

Here are the data from the *PLOTS* data file in the form needed by SYSTAT:

PLOT	A	B	YIELD
1	1	1	0
1	1	2	0
1	1	3	5
1	1	4	3
2	1	1	3
2	1	2	1
2	1	3	5
2	1	4	4
3	2	1	4
3	2	2	2
3	2	3	7
3	2	4	8
4	2	1	5
4	2	2	4
4	2	3	6
4	2	4	6

To analyze this design, you need two different error terms. For the between-plots effects (*A*), you need "plots within *A*". For the within-plots effects (*B* and *A*B*), you need "*B* by plots within *A*".

First, fit the saturated model with all the effects and then specify different error terms as needed.

The input is:

```
GLM
  USE PLOTS
  CATEGORY PLOT, A, B
  MODEL YIELD = CONSTANT + A + B + A*B + PLOT(A) + B*PLOT(A)
  ESTIMATE
```

The output is:

```
Dependent Variable : YIELD
N                  : 16
Multiple R         : 1.000
Squared Multiple R : 1.000
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
A	27.563	1	27.563	.	.
B	42.687	3	14.229	.	.
A*B	2.187	3	0.729	.	.
PLOT(A)	3.125	2	1.563	.	.
B*PLOT(A)	7.375	6	1.229	.	.
Error	0.000	0	.	.	.

You do not get a full ANOVA table because the model is perfectly fit. The coefficient of determination (squared multiple R) is 1. Now you have to use some of the effects as error terms.

Between-Plots Effects

Let's test for between-plots effects, namely *A*.

The input is:

```
HYPOTHESIS
EFFECT A
ERROR PLOT(A)
TEST
```

The output is:

Test for effect called: A

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A	27.563	1	27.563	17.640	0.052
Error	3.125	2	1.563		

The between-plots effect is not significant ($p\text{-value} = 0.052$).

Within-Plots Effects

To do the within-plots effects (*B* and *A*B*), the input is:

```
HYPOTHESIS
EFFECT B
ERROR B*PLOT(A)
TEST
HYPOTHESIS
EFFECT A*B
ERROR B*PLOT(A)
TEST
```

The output is:

Test for effect called: B

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	4.688	1	4.688	3.814	0.099
A2	25.521	1	25.521	20.763	0.004
A3	17.521	1	17.521	14.254	0.009
A	42.687	3	14.229	11.576	0.007
Error	7.375	6	1.229		

Test for effect called: A*B

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	0.188	1	0.188	0.153	0.710
A2	0.021	1	0.021	0.017	0.901
A3	1.688	1	1.688	1.373	0.286
A	2.187	3	0.729	0.593	0.642
Error	7.375	6	1.229		

Here, we find a significant effect due to factor *B* ($p\text{-value} = 0.007$), but the interaction is not significant ($p\text{-value} = 0.642$).

This analysis is the same as that for a repeated-measures design with subjects as *PLOT*, groups as *A*, and trials as *B*. Because this method becomes unwieldy for a large number of plots (subjects), SYSTAT offers a more compact method for repeated measures analysis as an alternative.

Example 8

Latin Square Designs

A Latin square design imposes a pattern on treatments in a factorial design to save experimental effort or reduce within-cell error. As in the nested design, not all combinations of the square and other treatments are measured, so the model lacks certain interaction terms between squares and treatments. GLM can analyze these designs easily if an extra variable denoting the square is included in the file. The following fixed-effects example is from Kutner et al. (2004). The *SQUARE* variable is

represented in the cells of the design. For simplicity, the dependent variable, *RESPONSE*, has been left out.

	day1	day2	day3	day4	day5
week1	D	C	A	B	E
week2	C	B	E	A	D
week3	A	D	B	E	C
week4	E	A	C	D	B
week5	B	E	D	C	A

You would set up the data as shown below (the *LATIN* file).

DAY	WEEK	SQUARE	RESPONSE
1	1	D	18
1	2	C	17
1	3	A	14
1	4	E	21
1	5	B	17
2	1	C	13
2	2	B	34
2	3	D	21
2	4	A	16
2	5	E	15
3	1	A	7
3	2	E	29
3	3	B	32
3	4	C	27
3	5	D	13
4	1	B	17
4	2	A	13
4	3	E	24
4	4	D	31
4	5	C	25
5	1	E	21
5	2	D	26
5	3	C	26
5	4	B	31
5	5	A	7

To do the analysis, the input is:

```
GLM
USE LATIN
CATEGORY DAY, WEEK, SQUARE
MODEL RESPONSE = CONSTANT + DAY + WEEK + SQUARE
ESTIMATE
```

The output is:

```
Dependent Variable : RESPONSE
N : 25
Multiple R : 0.931
Squared Multiple R : 0.867
```

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
DAY	82.000	4	20.500	1.306	0.323
WEEK	477.200	4	119.300	7.599	0.003
SQUARE	664.400	4	166.100	10.580	0.001
Error	188.400	12	15.700		

Example 9

Crossover and Changeover Designs

In crossover designs, an experiment is divided into periods, and the treatment of a subject changes from one period to the next. Changeover studies often use designs similar to a Latin square. A problem with these designs is that there may be a residual or carry-over effect of a treatment into the following period. This can be minimized by extending the interval between experimental periods; however, this is not always feasible. Fortunately, there are methods to assess the magnitude of any carry-over effects that may be present.

Two-period crossover designs can be analyzed as repeated-measures designs. More complicated crossover designs can also be analyzed by SYSTAT, and carry-over effects can be assessed. Cochran and Cox (1957) present a study of milk production by cows under three different feed schedules: *A* (roughage), *B* (limited grain), and *C* (full grain). The design of the study has the form of two (3×3) Latin squares:

COW						
Latin square 1				Latin square 2		
Period	I	II	III	IV	V	VI
1	A	B	C	A	B	C
2	B	C	A	C	A	B
3	C	A	B	B	C	A

The data are set up in the *WILLIAMS* data file as follows:

COW	SQUARE	PERIOD	FEED	CARRY	RESIDUAL	MILK
1	1	1	1	1	0	38
1	1	2	2	1	1	25
1	1	3	3	2	2	15
2	1	1	2	1	0	109
2	1	2	3	2	2	86
2	1	3	1	2	3	39
3	1	1	3	1	0	124
3	1	2	1	2	3	72
3	1	3	2	1	1	27
4	2	1	1	1	0	86
4	2	2	3	1	1	76
4	2	3	2	2	3	46
5	2	1	2	1	0	75
5	2	2	1	2	2	35
5	2	3	3	1	1	34
6	2	1	3	1	0	101
6	2	2	2	2	3	63
6	2	3	1	2	2	1

PERIOD is nested within each Latin square (the periods for cows in one square are unrelated to the periods in the other). The variable *RESIDUAL* indicates the treatment of the preceding period. For the first period for each cow, there is no preceding period. The input is:

```
GLM
  USE WILLIAMS
  CATEGORY COW, PERIOD, SQUARE, RESIDUAL, CARRY, FEED
  MODEL MILK = CONSTANT+COW+FEED+PERIOD (SQUARE)+RESIDUAL (CARRY)
  ESTIMATE
```

The output is:

Dependent Variable	MILK
N	18
Multiple R	0.995
Squared Multiple R	0.990

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
COW	3835.950	5	767.190	15.402	0.010
FEED	2854.550	2	1427.275	28.653	0.004
PERIOD(SQUARE)	3873.950	4	968.488	19.443	0.007
RESIDUAL(CARRY)	616.194	2	308.097	6.185	0.060
Error	199.250	4	49.813		

There is a significant effect of feed on milk production and an insignificant residual or carry-over effect in this instance.

Type I Sum-of-Squares Analysis

To replicate the Cochran and Cox Type I sum-of-squares analysis, you must fit a new model to get their sum of squares.

The input is:

```
GLM
CATEGORY COW
MODEL MILK = CONSTANT + COW + FEED +
PERIOD(SQUARE)+RESIDUAL(CARRY)
ESTIMATE / SS = TYPE1
```

The output is:

```
Dependent Variable   MILK
N                     18
Multiple R            0.995
Squared Multiple R    0.990
```

Analysis of Variance

Source	Type I SS	df	Mean Squares	F-ratio	p-value
COW	5781.111	5	1156.222	23.211	0.005
FEED	2276.778	2	1138.389	22.853	0.006
PERIOD(SQUARE)	11489.111	4	2872.278	57.662	0.001
RESIDUAL(CARRY)	616.194	2	308.097	6.185	0.060
Error	199.250	4	49.813		

Example 10**Missing Cells Designs (the Means Model)**

When cells are completely missing in a factorial design, parameterizing a model can be difficult. The full model cannot be estimated. GLM offers a means model parameterization so that missing cell parameters can be dropped automatically from the model, and hypotheses for main effects and interactions can be tested by specifying

cells directly. Examine Searle (1987), Hocking (1985), or Milliken and Johnson (1984) for more information in this area.

Widely favored for this purpose by statisticians (Searle, 1987; Hocking, 1985; Milliken and Johnson, 1984), the means model allows:

- Tests of hypotheses in missing cells designs (using Type IV sum of squares)
- Tests of simple hypotheses (for example, within levels of other factors)
- The use of population weights to reflect differences in subclass sizes

Effects coding is the default for GLM. Alternatively, means models code predictors as cell means rather than effects, which differ from a grand mean. The constant is omitted, and the predictors are 1 for a case belonging to a given cell and 0 for all others. When cells are missing, GLM automatically excludes null columns and estimates the submodel.

The categorical variables are specified in the MODEL statement differently for a means model than for an effects model. Here are some examples:

```
MODEL Y = A*B / MEANS
```

```
MODEL Y = GROUP*AGE*SCHOOL$ / MEANS
```

These two models generate fully factorial designs (A by B and group by AGE by SCHOOL\$). Notice that they omit the constant and main effects parameters because the means model does not include effects or a grand mean. Nevertheless, the number of parameters is the same in the two models. The following are the effects model and the means model, respectively, for a 2×3 design (two levels of A and three levels of B):

```
MODEL Y = CONSTANT + A + B + A*B
```

A	B	m	a1	b1	b2	a1b1	a1b2
1	1	1	1	1	0	1	0
1	2	1	1	0	1	0	1
1	3	1	1	-1	-1	-1	-1
2	1	1	-1	1	0	-1	0
2	2	1	-1	0	1	0	-1
2	3	1	-1	-1	-1	1	-1

```
MODEL Y = A*B / MEANS
```


A	B	a1b1	a1b2	a1b3	a2b1	a2b2	a2b3
1	1	1	0	0	0	0	0
1	2	0	1	0	0	0	0
1	3	0	0	1	0	0	0
2	1	0	0	0	1	0	0
2	2	0	0	0	0	1	0
2	3	0	0	0	0	0	1

Means and effects models can be blended for incomplete factorials and others designs. All crossed terms (for example, $A*B$) will be coded with means design variables (provided the MEANS option is present), and the remaining terms will be coded as effects. The constant must be omitted, even in these cases, because it is collinear with the means design variables. All covariates and effects that are coded factors must precede the crossed factors in the MODEL statement.

Here is an example, assuming A has four levels, B has two, and C has three. In this design, there are 24 possible cells, but only 12 are nonmissing. The treatment combinations are partially balanced across the levels of B and C.

MODEL Y = A + B*C / MEANS

A	B	C	a1	a2	a3	b1c1	b1c2	b1c3	b2c1	b2c2	b2c3
1	1	1	1	0	0	1	0	0	0	0	0
3	1	1	0	0	1	1	0	0	0	0	0
2	1	2	0	1	0	0	1	0	0	0	0
4	1	2	-1	-1	-1	0	1	0	0	0	0
1	1	3	1	0	0	0	0	1	0	0	0
4	1	3	-1	-1	-1	0	0	1	0	0	0
2	2	1	0	1	0	0	0	0	1	0	0
3	2	1	0	0	1	0	0	0	0	1	0
2	2	2	0	1	0	0	0	0	0	1	0
4	2	2	-1	-1	-1	0	0	0	0	1	0
1	2	3	1	0	0	0	0	0	0	0	1
3	2	3	0	0	1	0	0	0	0	0	1

Nutritional Knowledge Survey

The following example, which uses the data file *MJ202*, is from Milliken and Johnson (1984). The data are from a home economics survey experiment. *DIFF* is the change in test scores between pre-test and post-test on a nutritional knowledge questionnaire. *GROUP* classifies whether or not a subject received food stamps. *AGE* designates four age groups, and *RACE\$* was their term for designating Whites, Blacks, and Hispanics.

	Group 0				Group 1			
	1	2	3	4	1	2	3	4
W	1	3	6		9	10	13	15
H			5				12	
B		2	4	7	8		11	14

Empty cells denote age/race combinations for which no data were collected. Numbers within cells refer to cell designations in the Fisher LSD pairwise mean comparisons at the end of this example.

To first fit the model, the input is:

```
GLM
  USE MJ202
  CATEGORY GROUP AGE RACE$
  MODEL DIFF = GROUP*AGE*RACE$ / MEANS
  ESTIMATE
```

The output is:

Means Model

Dependent Variable	DIFF
N	107
Multiple R	0.538
Squared Multiple R	0.289

*** WARNING *** : Missing cells encountered. Tests of factors will not appear.

Ho: All means equal.
Unweighted Means Model

Analysis of Variance

Source	SS	df	Mean Squares	F-ratio	p-value
Model	1068.546	14	76.325	2.672	0.003
Error	2627.472	92	28.559		

We need to test the *GROUP* main effect. The following notation is equivalent to Milliken and Johnson's. Because of the missing cells, the *GROUP* effect must be computed over means that are balanced across the other factors.

In the drawing at the beginning of this example, notice that this specification contrasts all the numbered cells in group 0 (except 2) with all the numbered cells in group 1 (except 8 and 15).

The input is:

```

HYPOTHESIS
NOTE 'GROUP MAIN EFFECT
SPECIFY
    GROUP [0] AGE [1] RACE$ [W] + GROUP [0] AGE [2] RACE$ [W] +,
    GROUP [0] AGE [3] RACE$ [B] + GROUP [0] AGE [3] RACE$ [H] +,
    GROUP [0] AGE [3] RACE$ [W] + GROUP [0] AGE [4] RACE$ [B] =,
    GROUP [1] AGE [1] RACE$ [W] + GROUP [1] AGE [2] RACE$ [W] +,
    GROUP [1] AGE [3] RACE$ [B] + GROUP [1] AGE [3] RACE$ [H] +,
    GROUP [1] AGE [3] RACE$ [W] + GROUP [1] AGE [4] RACE$ [B]
TEST

```

The output is:

A Matrix

1	2	3	4	5
1.000	0.000	1.000	1.000	1.000

A Matrix

6	7	8	9	10
1.000	1.000	0.000	-1.000	-1.000

A Matrix

11	12	13	14	15
-1.000	-1.000	-1.000	-1.000	0.000

Null Hypothesis Value for D

0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
Hypothesis	75.738	1	75.738	2.652	0.107
Error	2627.472	92	28.559		

The computations for the *AGE* main effect are similar to those for the *GROUP* main effect:

HYPOTHESIS

NOTE 'AGE MAIN EFFECT'

SPECIFY,

```

GROUP [1] AGE [1] RACE$ [B] + GROUP [1] AGE [1] RACE$ [W] =,
GROUP [1] AGE [4] RACE$ [B] + GROUP [1] AGE [4] RACE$ [W] ;,
GROUP [0] AGE [2] RACE$ [B] + GROUP [1] AGE [2] RACE$ [W] =,
GROUP [0] AGE [4] RACE$ [B] + GROUP [1] AGE [4] RACE$ [W] ;,
GROUP [0] AGE [3] RACE$ [B] + GROUP [1] AGE [3] RACE$ [B] +,
GROUP [1] AGE [3] RACE$ [W] =,
GROUP [0] AGE [4] RACE$ [B] + GROUP [1] AGE [4] RACE$ [B] +,
GROUP [1] AGE [4] RACE$ [W]

```

TEST

The output is:

A Matrix

	1	2	3	4	5
1	0.000	0.000	0.000	0.000	0.000
2	0.000	1.000	0.000	0.000	0.000
3	0.000	0.000	0.000	1.000	0.000

A Matrix

	6	7	8	9	10
1	0.000	0.000	1.000	1.000	0.000
2	0.000	-1.000	0.000	0.000	1.000
3	0.000	-1.000	0.000	0.000	0.000

A Matrix

	11	12	13	14	15
1	0.000	0.000	0.000	-1.000	-1.000
2	0.000	0.000	0.000	0.000	-1.000
3	1.000	0.000	1.000	-1.000	-1.000

D Matrix

1	0.000
2	0.000
3	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	7.139	1	7.139	0.250	0.618
A2	0.202	1	0.202	0.007	0.933
A3	29.231	1	29.231	1.024	0.314
A	41.526	3	13.842	0.485	0.694
Error	2627.472	92	28.559		

The *GROUP* by *AGE* interaction requires more complex balancing than the main effects. It is derived from a subset of the means in the following specified combination. Again, check Milliken and Johnson to see the correspondence.

The input is:

HYPOTHESIS

NOTE 'GROUP BY AGE INTERACTION'

SPECIFY,

```

GROUP[0] AGE[1] RACE$[W] - GROUP[0] AGE[3] RACE$[W] -,
GROUP[1] AGE[1] RACE$[W] + GROUP[1] AGE[3] RACE$[W] +,
GROUP[0] AGE[3] RACE$[B] - GROUP[0] AGE[4] RACE$[B] -,
GROUP[1] AGE[3] RACE$[B] + GROUP[1] AGE[4] RACE$[B]=0.0;,
GROUP[0] AGE[2] RACE$[W] - GROUP[0] AGE[3] RACE$[W] -,
GROUP[1] AGE[2] RACE$[W] + GROUP[1] AGE[3] RACE$[W] +,
GROUP[0] AGE[3] RACE$[B] - GROUP[0] AGE[4] RACE$[B] -,
GROUP[1] AGE[3] RACE$[B] + GROUP[1] AGE[4] RACE$[B]=0.0;,
GROUP[0] AGE[3] RACE$[B] - GROUP[0] AGE[4] RACE$[B] -,
GROUP[1] AGE[3] RACE$[B] + GROUP[1] AGE[4] RACE$[B]=0.0

```

TEST

The output is:

A Matrix

	1	2	3	4	5
1	1.000	0.000	0.000	1.000	0.000
2	0.000	0.000	1.000	1.000	0.000
3	0.000	0.000	0.000	1.000	0.000

A Matrix

	6	7	8	9	10
1	-1.000	-1.000	0.000	-1.000	0.000
2	-1.000	-1.000	0.000	0.000	-1.000
3	0.000	-1.000	0.000	0.000	0.000

A Matrix

	11	12	13	14	15
1	-1.000	0.000	1.000	1.000	0.000
2	-1.000	0.000	1.000	1.000	0.000
3	-1.000	0.000	0.000	1.000	0.000

D Matrix

1	0.000
2	0.000
3	0.000

Test of Hypothesis

Source	SS	df	Mean Squares	F-ratio	p-value
A1	38.868	1	38.868	1.361	0.246
A2	71.783	1	71.783	2.513	0.116
A3	83.046	1	83.046	2.908	0.092
A	91.576	3	30.525	1.069	0.366
Error	2627.472	92	28.559		

The following commands are needed to produce the rest of Milliken and Johnson's results. The remaining output is not listed.

```

HYPOTHESIS
NOTE 'RACE$ MAIN EFFECT'
SPECIFY,
    GROUP[0] AGE[2] RACE$[B] + GROUP[0] AGE[3] RACE$[B] +,
    GROUP[1] AGE[1] RACE$[B] + GROUP[1] AGE[3] RACE$[B] +,
    GROUP[1] AGE[4] RACE$[B] =,
    GROUP[0] AGE[2] RACE$[W] + GROUP[0] AGE[3] RACE$[W] +,
    GROUP[1] AGE[1] RACE$[W] + GROUP[1] AGE[3] RACE$[W] +,
    GROUP[1] AGE[4] RACE$[W] ;,
    GROUP[0] AGE[3] RACE$[H] + GROUP[1] AGE[3] RACE$[H] =,
    GROUP[0] AGE[3] RACE$[W] + GROUP[1] AGE[3] RACE$[W]

TEST
HYPOTHESIS
NOTE 'GROUP*RACE$'
SPECIFY,
    GROUP[0] AGE[3] RACE$[B] - GROUP[0] AGE[3] RACE$[W] -,
    GROUP[1] AGE[3] RACE$[B] + GROUP[1] AGE[3]
    RACE$[W]=0.0;;
    GROUP[0] AGE[3] RACE$[H] - GROUP[0] AGE[3] RACE$[W] -,
    GROUP[1] AGE[3] RACE$[H] + GROUP[1] AGE[3] RACE$[W]=0.0

TEST
HYPOTHESIS
NOTE 'AGE*RACE$'
SPECIFY,
    GROUP[1] AGE[1] RACE$[B] - GROUP[1] AGE[1] RACE$[W] -,
    GROUP[1] AGE[4] RACE$[B] + GROUP[1] AGE[4]
    RACE$[W]=0.0;;
    GROUP[0] AGE[2] RACE$[B] - GROUP[0] AGE[2] RACE$[W] -,
    GROUP[0] AGE[3] RACE$[B] + GROUP[0] AGE[3]
    RACE$[W]=0.0;;
    GROUP[1] AGE[3] RACE$[B] - GROUP[1] AGE[3] RACE$[W] -,
    GROUP[1] AGE[4] RACE$[B] + GROUP[1] AGE[4] RACE$[W]=0.0

TEST

```

Finally, Milliken and Johnson do pairwise comparisons:

```

HYPOTHESIS
POST GROUP*AGE*RACE$ / LSD
TEST

```

The following is the matrix of comparisons printed by GLM. The matrix of mean differences has been omitted.

Post Hoc Test of DIFF
Using unweighted means.
Using model MSE of 28.559 with 92 df.

Fisher's Least-Significant-Difference Test

GROUP (i) * AGE (i-)* RACE\$ (i)	GROUP (j) * AGE (j-)* RACE\$ (j)	Difference	p-value	95.0% Confidence Interval	
				Lower	Upper
0*1*W	0*2*B	-1.619	0.662	-8.943	5.705
0*1*W	0*2*W	-1.708	0.638	-8.894	5.477
0*1*W	0*3*B	1.333	0.725	-6.172	8.838
0*1*W	0*3*H	-4.833	0.324	-14.522	4.856
0*1*W	0*3*W	-2.133	0.521	-8.705	4.438
0*1*W	0*4*B	-2.333	0.706	-14.589	9.922
0*1*W	1*1*B	-6.333	0.197	-16.022	3.356
0*1*W	1*1*W	-2.833	0.563	-12.522	6.856
0*1*W	1*2*W	-7.208	0.049	-14.394	-0.023
0*1*W	1*3*B	-9.833	0.018	-17.940	-1.727
0*1*W	1*3*H	-2.333	0.706	-14.589	9.922
0*1*W	1*3*W	-7.753	0.018	-14.170	-1.335
0*1*W	1*4*B	0.667	0.914	-11.589	12.922
0*1*W	1*4*W	-5.970	0.090	-12.883	0.944
0*2*B	0*2*W	-0.089	0.974	-5.582	5.404
0*2*B	0*3*B	2.952	0.323	-2.953	8.857
0*2*B	0*3*H	-3.214	0.455	-11.724	5.296
0*2*B	0*3*W	-0.514	0.827	-5.175	4.147
0*2*B	0*4*B	-0.714	0.901	-12.061	10.632
0*2*B	1*1*B	-4.714	0.274	-13.224	3.796
0*2*B	1*1*W	-1.214	0.778	-9.724	7.296
0*2*B	1*2*W	-5.589	0.046	-11.082	-0.096
0*2*B	1*3*B	-8.214	0.016	-14.867	-1.562
0*2*B	1*3*H	-0.714	0.901	-12.061	10.632
0*2*B	1*3*W	-6.134	0.007	-10.575	-1.692
0*2*B	1*4*B	2.286	0.690	-9.061	13.632
0*2*B	1*4*W	-4.351	0.096	-9.482	0.781
0*2*W	0*3*B	3.042	0.295	-2.690	8.774
0*2*W	0*3*H	-3.125	0.461	-11.516	5.266
0*2*W	0*3*W	-0.425	0.850	-4.865	4.015
0*2*W	0*4*B	-0.625	0.912	-11.883	10.633
0*2*W	1*1*B	-4.625	0.277	-13.016	3.766
0*2*W	1*1*W	-1.125	0.791	-9.516	7.266
0*2*W	1*2*W	-5.500	0.042	-10.807	-0.193
0*2*W	1*3*B	-8.125	0.015	-14.625	-1.625
0*2*W	1*3*H	-0.625	0.912	-11.883	10.633
0*2*W	1*3*W	-6.044	0.005	-10.253	-1.835
0*2*W	1*4*B	2.375	0.676	-8.883	13.633
0*2*W	1*4*W	-4.261	0.090	-9.193	0.670
0*3*B	0*3*H	-6.167	0.161	-14.833	2.500
0*3*B	0*3*W	-3.467	0.167	-8.407	1.474
0*3*B	0*4*B	-3.667	0.527	-15.131	7.798
0*3*B	1*1*B	-7.667	0.082	-16.333	1.000
0*3*B	1*1*W	-4.167	0.342	-12.833	4.500
0*3*B	1*2*W	-8.542	0.004	-14.274	-2.810
0*3*B	1*3*B	-11.167	0.002	-18.018	-4.315
0*3*B	1*3*H	-3.667	0.527	-15.131	7.798
0*3*B	1*3*W	-9.086	0.000	-13.820	-4.352
0*3*B	1*4*B	-0.667	0.908	-12.131	10.798
0*3*B	1*4*W	-7.303	0.008	-12.690	-1.916
0*3*H	0*3*W	2.700	0.497	-5.171	10.571
0*3*H	0*4*B	2.500	0.703	-10.499	15.499
0*3*H	1*1*B	-1.500	0.780	-12.114	9.114
0*3*H	1*1*W	2.000	0.709	-8.614	12.614
0*3*H	1*2*W	-2.375	0.575	-10.766	6.016
0*3*H	1*3*B	-5.000	0.283	-14.192	4.192

Linear Models III: General Linear Models

0*3*H	1*3*H	2.500	0.703	-10.499	15.499
0*3*H	1*3*W	-2.919	0.456	-10.663	4.824
0*3*H	1*4*B	5.500	0.403	-7.499	18.499
0*3*H	1*4*W	-1.136	0.783	-9.295	7.023
0*3*W	0*4*B	-0.200	0.971	-11.076	10.676
0*3*W	1*1*B	-4.200	0.292	-12.071	3.671
0*3*W	1*1*W	-0.700	0.860	-8.571	7.171
0*3*W	1*2*W	-5.075	0.026	-9.515	-0.635
0*3*W	1*3*B	-7.700	0.010	-13.513	-1.887
0*3*W	1*3*H	-0.200	0.971	-11.076	10.676
0*3*W	1*3*W	-5.619	0.000	-8.663	-2.575
0*3*W	1*4*B	2.800	0.610	-8.076	13.676
0*3*W	1*4*W	-3.836	0.059	-7.821	0.148
0*4*B	1*1*B	-4.000	0.543	-16.999	8.999
0*4*B	1*1*W	-0.500	0.939	-13.499	12.499
0*4*B	1*2*W	-4.875	0.392	-16.133	6.383
0*4*B	1*3*B	-7.500	0.213	-19.367	4.367
0*4*B	1*3*H	0.000	1.000	-15.010	15.010
0*4*B	1*3*W	-5.419	0.321	-16.203	5.364
0*4*B	1*4*B	3.000	0.692	-12.010	18.010
0*4*B	1*4*W	-3.636	0.516	-14.722	7.449
1*1*B	1*1*W	3.500	0.514	-7.114	14.114
1*1*B	1*2*W	-0.875	0.836	-9.266	7.516
1*1*B	1*3*B	-3.500	0.451	-12.692	5.692
1*1*B	1*3*H	4.000	0.543	-8.999	16.999
1*1*B	1*3*W	-1.419	0.717	-9.163	6.324
1*1*B	1*4*B	7.000	0.288	-5.999	19.999
1*1*B	1*4*W	0.364	0.930	-7.795	8.523
1*1*W	1*2*W	-4.375	0.303	-12.766	4.016
1*1*W	1*3*B	-7.000	0.134	-16.192	2.192
1*1*W	1*3*H	0.500	0.939	-12.499	13.499
1*1*W	1*3*W	-4.919	0.210	-12.663	2.624
1*1*W	1*4*B	3.500	0.594	-9.499	16.499
1*1*W	1*4*W	-3.136	0.447	-11.295	5.023
1*2*W	1*3*B	-2.625	0.425	-9.125	3.875
1*2*W	1*3*H	4.875	0.392	-6.383	16.133
1*2*W	1*3*W	-0.544	0.798	-4.753	3.665
1*2*W	1*4*B	7.875	0.168	-3.383	19.133
1*2*W	1*4*W	1.239	0.619	-3.693	6.170
1*3*B	1*3*H	7.500	0.213	-4.367	19.367
1*3*B	1*3*W	2.081	0.466	-3.558	7.720
1*3*B	1*4*B	10.500	0.082	-1.367	22.367
1*3*B	1*4*W	3.864	0.219	-2.334	10.061
1*3*H	1*4*W	-5.419	0.321	-16.203	5.364
1*3*H	1*3*W	3.000	0.692	-12.010	18.010
1*3*H	1*4*B	-3.636	0.516	-14.722	7.449
1*3*H	1*4*W	8.419	0.124	-2.364	19.203
1*3*W	1*4*B	1.783	0.344	-1.942	5.508
1*3*W	1*4*W	-6.636	0.238	-17.722	4.449

* This test controls the comparisonwise error rate but not the family-wise error rate.

Within group 0 (cells 1–7), there are no significant pairwise differences in the average test score changes. The same is true within group 1 (cells 8–15).

Example 11

Covariance Alternatives to Repeated Measures

Analysis of covariance offers an alternative to repeated measures in a pre-post design. You can use the pre-test as a covariate in predicting the post-test. This example shows how to do a two-group, pre-post design:

```
GLM
  USE FILENAME
  CATEGORY GROUP
  MODEL POST = CONSTANT + GROUP + PRE
  ESTIMATE
```

When using this design, be sure to check the homogeneity of slopes assumption. Use the following commands to check that the interaction term, GROUP*PRE, is not significant:

```
GLM
  USE FILENAME
  CATEGORY GROUP
  MODEL POST = CONSTANT + GROUP + PRE + GROUP*PRE
  ESTIMATE
```

Example 12

Weighting Means

Sometimes you want to weight the cell means when you test hypotheses in ANOVA. Suppose you have an experiment in which a few rats died before its completion. You do not want the hypotheses tested to depend upon the differences in cell sizes (which are presumably random). Here is an example from Morrison (2004). The data (*MOTHERS*) are hypothetical profiles on three scales of mothers in each of four socioeconomic classes.

Morrison analyzes these data with the multivariate profile model for repeated measures. Because the hypothesis of parallel profiles across classes is not rejected, you can test whether the profiles are level. That is, do the scales differ when we pool the classes together?

Pooling unequal classes can be done by weighting each according to sample size or averaging the means of the subclasses. First, let's look at the model and test the hypothesis of equality of scale parameters without weighting the cell means.

The input is:

```
GLM
USE MOTHERS
MODEL SCALE(1) SCALE(2) SCALE(3) = CONSTANT+CLASS
CATEGORY CLASS / EFFECT
ESTIMATE
HYPOTHESIS
EFFECT CONSTANT
CMATRIX [ 1 -1 0; 0 1 -1 ]
TEST
```

The output is:

Dependent Variable Means

SCALE(1)	SCALE(2)	SCALE(3)
14.524	15.619	15.857

Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	SCALE(1)	SCALE(2)	SCALE(3)
CONSTANT		13.700	14.550	14.988
CLASS	1	4.300	5.450	4.763
CLASS	2	0.100	0.650	-0.787
CLASS	3	-0.700	-0.550	0.012

Test for effect called: CONSTANT

C Matrix

	1	2	3
1	1.000	-1.000	0.000
2	0.000	1.000	-1.000

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
1	14.012	1	14.012	4.652	0.046
Error	51.200	17	3.012		
2	3.712	1	3.712	1.026	0.325
Error	61.500	17	3.618		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.564	6.191	2, 16	0.010
Pillai Trace	0.436	6.191	2, 16	0.010
Hotelling-Lawley Trace	0.774	6.191	2, 16	0.010

Notice that the dependent variable means differ from the *CONSTANT*. The *CONSTANT* in this case is a mean of the cell means rather than the mean of all the cases.

Weighting by the Sample Size

Suppose you believe (as Morrison does) that the differences in cell sizes reflect population subclass proportions, then you need to weight the cell means to get a grand mean; for example:

$$8(\mu_1) + 5(\mu_2) + 4(\mu_3) + 4(\mu_4)$$

Expressed in terms of our analysis of variance parameterization, this is:

$$8(\mu + \alpha_1) + 5(\mu + \alpha_2) + 4(\mu + \alpha_3) + 4(\mu + \alpha_4)$$

Because the sum of effects is 0 for a classification and because you do not have an independent estimate of *CLASS4*, this expression is equivalent to:

$$8(\mu + \alpha_1) + 5(\mu + \alpha_2) + 4(\mu + \alpha_3) + 4(\mu - \alpha_1 - \alpha_2 - \alpha_3)$$

which works out to:

$$21\mu + 4(\alpha_1) + 1(\alpha_2) + 0(\alpha_3)$$

Use AMATRIX to test this hypothesis.

The input is:

```
HYPOTHESIS
AMATRIX [21  4  1  0]
CMATRIX [1 -1  0; 0  1 -1]
TEST
```

The output is:

A Matrix

	1	2	3	4
21.000	4.000	1.000	0.000	

C Matrix

	1	2	3
1	1.000	-1.000	0.000
2	0.000	1.000	-1.000

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
1	25.190	1	25.190	8.364	0.010
Error	51.200	17	3.012		
2	1.190	1	1.190	0.329	0.574
Error	61.500	17	3.618		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.501	7.959	2, 16	0.004
Pillai Trace	0.499	7.959	2, 16	0.004
Hotelling-Lawley Trace	0.995	7.959	2, 16	0.004

This is the multivariate *F-ratio* statistic that Morrison gets. For these data, we prefer the weighted means analysis because these differences in cell frequencies probably reflect population base rates. They are not random.

Example 13

Hotelling's T-Square

You can use GLM to calculate Hotelling's T-square statistic.

One-Sample Test

For example, to get a one-sample test for the variables *X* and *Y*, select both *X* and *Y* as dependent variables.

The input is:

```
GLM
USE FILENAME
MODEL X, Y = CONSTANT
ESTIMATE
```

The F-test for *CONSTANT* is the statistic you want. It is the same as the Hotelling's T^2 for the hypothesis that the population means for *X* and *Y* are 0.

You can also test against the hypothesis that the means of *X* and *Y* have particular nonzero values (for example, 10 and 15) by using:

```
HYPOTHESIS
DMATRIX [10 15]
TEST
```


Two-Sample Test

For a two-sample test, you must provide a categorical independent variable that represents the two groups.

The input is:

```
GLM
CATEGORY GROUP
MODEL X, Y = CONSTANT + GROUP
ESTIMATE
```

Example 14 Discriminant Analysis

This example uses the *IRIS* data file. Fisher used these data to illustrate his discriminant function. To define the model:

```
GLM
USE IRIS
CATEGORY SPECIES
MODEL SEPALLEN  SEPALWID  PETALLEN  PETALWID = CONSTANT +,
                                SPECIES
ESTIMATE
HYPOTHESIS
EFFECT SPECIES
SAVE CANON
TEST
```

SYSTAT saves the canonical scores associated with the hypothesis. The scores are stored in subscripted variables named *FACTOR*. Because the effects involve a categorical variable, the Mahalanobis distances (named *DISTANCE*) and posterior probabilities (named *PROB*) are saved in the same file. These distances are computed in the discriminant space itself. The closer a case is to a particular group's location in that space, the more likely it is that it belongs to that group. The probability of group membership is computed from these distances. A variable named *PREDICT* that contains the predicted group membership is also added to the file.

The output is:

Dependent Variable Means

SEPALLEN	SEPALWID	PETALLEN	PETALWID
5.843	3.057	3.758	1.199

Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	SEPALLEN	SEPALWID	PETALLEN	PETALWID
CONSTANT		5.843	3.057	3.758	1.199
SPECIES	1	-0.837	0.371	-2.296	-0.953
SPECIES	2	0.093	-0.287	0.502	0.127

Test for effect called: SPECIES

Null Hypothesis Contrast AB

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
1	-0.837	0.371	-2.296	-0.953
2	0.093	-0.287	0.502	0.127

Inverse Contrast $A(X'X)^{-1}A'$

	1	2
1	0.013	
2	-0.007	0.013

Hypothesis Sum of Product Matrix $H = B'A'(A(X'X)^{-1}A')^{-1}AB$

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	63.212			
SEPALWID	-19.953	11.345		
PETALLEN	165.248	-57.240	437.103	
PETALWID	71.279	-22.933	186.774	80.413

Error Sum of Product Matrix $G = E'E$

	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	38.956			
SEPALWID	13.630	16.962		
PETALLEN	24.625	8.121	27.223	
PETALWID	5.645	4.808	6.272	6.157

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
SEPALLEN	63.212	2	31.606	119.265	0.000
Error	38.956	147	0.265		
SEPALWID	11.345	2	5.672	49.160	0.000
Error	16.962	147	0.115		
PETALLEN	437.103	2	218.551	1180.161	0.000
Error	27.223	147	0.185		
PETALWID	80.413	2	40.207	960.007	0.000
Error	6.157	147	0.042		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.023	199.145	8, 288	0.000
Pillai Trace	1.192	53.466	8, 290	0.000
Hotelling-Lawley Trace	32.477	580.532	8, 286	0.000

THETA	S	M	N	p-value
0.970	2	0.500	71.000	0.000

Test of Residual Roots

Roots	Chi-square	df
1 through 2	546.115	8
2 through 2	36.530	3

Canonical Correlations

1	2
0.985	0.471

Dependent Variable Canonical Coefficients Standardized
by Conditional (within Groups) Standard Deviations

	1	2
SEPALLEN	0.427	0.012
SEPALWID	0.521	0.735
PETALLEN	-0.947	-0.401
PETALWID	-0.575	0.581

Canonical Loadings (Correlations between Conditional
Dependent Variables and Dependent Canonical Factors)

	1	2
SEPALLEN	-0.223	0.311
SEPALWID	0.119	0.864
PETALLEN	-0.706	0.168
PETALWID	-0.633	0.737

Group Classification Function Coefficients

	1	2	3
SEPALLEN	23.544	15.698	12.446
SEPALWID	23.588	7.073	3.685
PETALLEN	-16.431	5.211	12.767
PETALWID	-17.398	6.434	21.079

Group Classification Constants

1	2	3
-86.308	-72.853	-104.368

Canonical scores have been saved.

The multivariate tests are all significant. The dependent variable canonical coefficients are used to produce discriminant scores. These coefficients are standardized by the within-groups standard deviations so you can compare their magnitude across variables with different scales. Because they are not raw coefficients, there is no need for a constant. The scores produced by these coefficients have an overall zero mean and a unit standard deviation within groups.

The group classification coefficients and constants comprise the Fisher discriminant functions for classifying the raw data. You can apply these coefficients to new data and assign each case to the group with the largest function value for that case.

Studying Saved Results

The *CANON* file that was just saved contains the canonical variable scores (*FACTOR(1)* and *FACTOR(2)*), the Mahalanobis distances to each group centroid (*DISTANCE(1)*, *DISTANCE(2)*, and *DISTANCE(3)*), the posterior probability for each case being assigned to each group (*PROB(1)*, *PROB(2)*, and *PROB(3)*), the predicted group membership (*PREDICT*), and the original group assignment (*GROUP*).

To produce a classification table of the group assignment against the predicted group membership and a plot of the second canonical variable against the first, the input is:

```
XTAB
USE CANON
PRINT NONE/ FREQ CHISQ
TABULATE GROUP * PREDICT
PLOT FACTOR(2)*FACTOR(1)/OVERLAY GROUP=GROUP COLOR=2,1,3,
      FILL=1,1,1 SYMBOL=4,8,5
```

The output is:

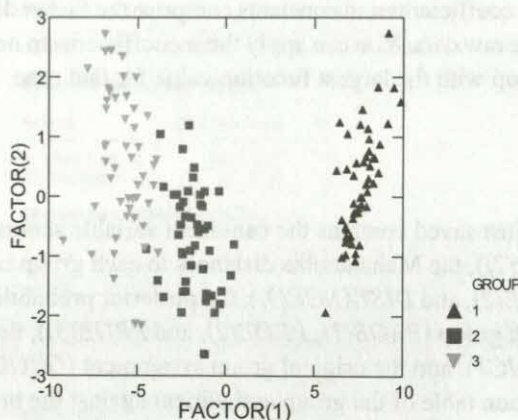
Counts

GROUP(rows) by PREDICT(columns)

	1	2	3	Total
1	50	0	0	50
2	0	48	2	50
3	0	1	49	50
Total	50	49	51	150

Chi-square Tests of Association for GROUP and PREDICT

Test Statistic	Value	df	p-value
Pearson Chi-square	282.593	4.000	0.000



However, it is much easier to use the Discriminant Analysis procedure.

Prior Probabilities

In this example, there were equal numbers of flowers in each group. Sometimes the probability of finding a case in each group is not the same across groups. To adjust the prior probabilities for this example, specify 0.5, 0.3, and 0.2 as the priors:

```
PRIORS 0.5 0.3 0.2
```

GLM uses the probabilities you specify to compute the posterior probabilities that are saved in the file under the variable *PROB*. Be sure to specify a probability for each level of the grouping variable. The probabilities should add up to 1.

Example 15

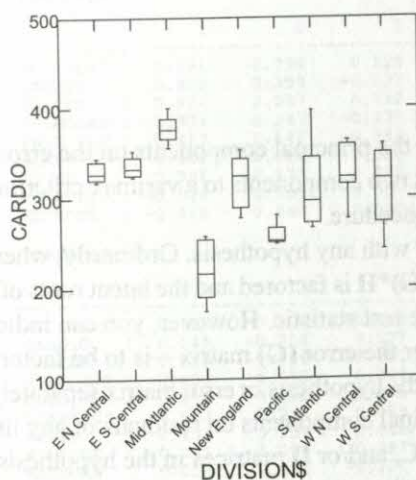
Principal Components Analysis (Within Groups)

GLM allows you to partial out effects based on grouping variables and to factor residual correlations. If between-group variation is significant, the within-group structure can differ substantially from the total structure (ignoring the grouping variable). However, if you are just computing principal components on a single sample (no grouping variable), you can obtain more detailed output using the Factor Analysis procedure.

The following data (*USSTATES*) comprise death rates by cause from nine census divisions of the country for that year. The divisions are in the column labeled *DIV*, and

the U.S. Post Office two-letter state abbreviations follow *DIV*. Other variables include *ACCIDENT*, *CARDIO*, *CANCER*, *PULMONAR*, *PNEU_FLU*, *DIABETES*, *LIVER*, *STATES*, *FSTROKE*, *MSTROKE*.

The variation in death rates between divisions in these data is substantial. Here is a grouped box plot of the second variable, *CARDIO*, by division. The other variables show similar regional differences.



Suppose you analyze these data ignoring *DIVISION\$*, the correlations among death rates would be due substantially to between-divisions differences. You might want to examine the pooled within-region correlations to see if the structure is different when divisional differences are statistically controlled. Accordingly, you will factor the residual correlation matrix after regressing medical variables onto an index variable denoting the census regions.

The input is:

```
GLM
USE USSTATES
CATEGORY DIVISION
MODEL ACCIDENT CARDIO CANCER PULMONAR PNEU FLU,
      DIABETES LIVER FSTROKE MSTROKE = CONSTANT + DIVISION
ESTIMATE
HYPOTHESIS
EFFECT DIVISION
FACTOR ERROR
TYPE CORR
ROTATE 2
TEST
```

The hypothesis commands compute the principal components on the error (residual) correlation matrix and rotate the first two components to a varimax criterion. For other rotations, use the Factor Analysis procedure.

The FACTOR options can be used with any hypothesis. Ordinarily, when you test a hypothesis, the matrix product $INV(G)*H$ is factored and the latent roots of this matrix are used to construct the multivariate test statistic. However, you can indicate which matrix—the hypothesis (**H**) matrix or the error (**G**) matrix—is to be factored. By computing principal components on the hypothesis or error matrix separately, FACTOR offers a direct way to compute principal components on residuals of any linear model you wish to fit. You can use any **A**, **C**, and/or **D** matrices in the hypothesis you are factoring, or you can use any of the other commands that create these matrices.

The output is:

Principal Components Computed on the following Error Correlation Matrix

	ACCIDENT	CARDIO	CANCER	PULMONAR	PNEU_FLU
ACCIDENT	1.000				
CARDIO	0.280	1.000			
CANCER	0.188	0.844	1.000		
PULMONAR	0.307	0.676	0.711	1.000	
PNEU FLU	0.113	0.448	0.297	0.396	1.000
DIABETES	0.297	0.419	0.526	0.296	-0.123
LIVER	-0.005	0.251	0.389	0.252	-0.138
FSTROKE	0.402	-0.202	-0.379	-0.190	-0.110
MSTROKE	0.495	-0.119	-0.246	-0.127	-0.071

Principal Components Computed on the following Error Correlation Matrix

	DIABETES	LIVER	FSTROKE	MSTROKE
DIABETES	1.000			
LIVER	-0.025	1.000		
FSTROKE	-0.151	-0.225	1.000	
MSTROKE	-0.076	-0.203	0.947	1.000

Latent Roots

1	2	3	4	5
3.341	2.245	1.204	0.999	0.475

Latent Roots

6	7	8	9
0.364	0.222	0.119	0.033

Loadings

	1	2	3	4	5
ACCIDENT	0.191	0.798	0.128	-0.018	-0.536
CARDIO	0.870	0.259	-0.097	0.019	0.219
CANCER	0.934	0.097	0.112	0.028	0.183
PULMONAR	0.802	0.247	-0.135	0.120	-0.071
PNEU FLU	0.417	0.146	-0.842	-0.010	-0.042
DIABETES	0.512	0.218	0.528	-0.580	0.068
LIVER	0.391	-0.175	0.400	0.777	-0.044
FSTROKE	-0.518	0.795	0.003	0.155	0.226
MSTROKE	-0.418	0.860	0.025	0.138	0.204

Loadings

	6	7	8	9
ACCIDENT	0.106	-0.100	-0.019	-0.015
CARDIO	0.145	-0.254	0.177	0.028
CANCER	0.039	-0.066	-0.251	-0.058
PULMONAR	-0.499	0.085	0.044	0.015
PNEU FLU	0.216	0.220	-0.005	-0.002
DIABETES	0.093	0.241	0.063	0.010
LIVER	0.154	0.159	0.046	0.009
FSTROKE	-0.041	0.056	0.081	-0.119
MSTROKE	0.005	0.035	-0.101	0.117

Rotated Loadings on first 2 Principal Components

	1	2
ACCIDENT	0.457	0.682
CARDIO	0.906	-0.060
CANCER	0.909	-0.234
PULMONAR	0.838	-0.047
PNEU FLU	0.441	-0.008
DIABETES	0.556	0.027
LIVER	0.305	-0.300
FSTROKE	-0.209	0.925
MSTROKE	-0.093	0.951

Sorted Rotated Loadings on first 2 Principal Components
(Loadings less than 0.25 made 0)

	1	2
ACCIDENT*CANCER	0.909	0.000
CARDIO*CARDIO	0.906	0.000
CANCER*PULMONAR	0.838	0.000
PULMONAR*DIABETES	0.556	0.000
PNEU FLU*MSTROKE	0.000	0.951
DIABETES*FSTROKE	0.000	0.925
LIVER*ACCIDENT	0.457	0.682
FSTROKE*LIVER	0.305	-0.300
MSTROKE*PNEU_FLU	0.441	0.000

Notice the sorted, rotated loadings. When interpreting these values, do not relate the row numbers (1 through 9) to the variables. Instead, find the corresponding loading in the Rotated Loadings table. The ordering of the rotated loadings corresponds to the order of the model variables.

The first component rotates to a dimension defined by *CANCER*, *CARDIO*, *PULMONAR*, and *DIABETES*; the second, by a dimension defined by *MSTROKE* and *FSTROKE* (male and female stroke rates). *ACCIDENT* also loads on the second factor but is not independent of the first. *LIVER* does not load highly on either factor.

Example 16 Canonical Correlation Analysis

Suppose you have 10 dependent variables, *MMPI(1)* to *MMPI(10)*, and 3 independent variables, *RATER(1)* to *RATER(3)*. Enter the following commands to obtain the canonical correlations and dependent canonical coefficients:

```
GLM
  USE DATAFILE
  MODEL MMPI(1 .. 10) = CONSTANT + RATER(1) + RATER(2) + RATER(3)
  ESTIMATE
  PLENGTH LONG
  HYPOTHESIS
  STANDARDIZE
  EFFECT RATER(1) & RATER(2) & RATER(3)
  TEST
```

The canonical correlations are displayed; you can rotate the dependent canonical coefficients by using the Rotate option.

To obtain the coefficients for the independent variables, run GLM again with the model reversed:

```
MODEL RATER(1 .. 3) = CONSTANT + MMPI(1) + MMPI(2),
                        + MMPI(3) + MMPI(4) + MMPI(5),
                        + MMPI(6) + MMPI(7) + MMPI(8),
                        + MMPI(9) + MMPI(10)

ESTIMATE
HYPOTHESIS
STANDARDIZE TOTAL
EFFECT MMPI(1) & MMPI(2) & MMPI(3) & MMPI(4) &, MMPI(5) &
        MMPI(6) & MMPI(7) & MMPI(8) &, MMPI(9) & MMPI(10)
TEST
```

Example 17

Mixture Models

Mixture models decompose the effects of mixtures of variables on a dependent variable. They differ from ordinary regression models because the independent variables sum to a constant value. The regression model, therefore, does not include a constant, and the regression and error sum of squares have one less degree of freedom. Marquardt and Snee (1974) and Diamond (2001) discuss these models and their estimation.

Here is an example using the *PUNCH* data file from Cornell (1985). The study involved effects of various mixtures of watermelon, pineapple, and orange juice on taste ratings by judges of a fruit punch.

The input is:

```
GLM
USE PUNCH
MODEL TASTE = WATRMELN + PINEAPPL + ORANGE + ,
               WATRMELN*PINEAPPL + WATRMELN*ORANGE + ,
               PINEAPPL*ORANGE
ESTIMATE / MIX
```

The output is:

Dependent Variable	TASTE
N	18
Multiple R	0.969
Squared Multiple R	0.939
Adjusted Squared Multiple R	0.913
Standard Error of Estimate	0.232

Regression Coefficients B = $(X'X)^{-1}X'Y$

Effect	Coefficient	Standard Error	Coefficient	Std. Tolerance	t	p-value
WATRMELN	4.600	0.134	3.001	0.667	34.322	0.000
PINEAPPL	6.333	0.134	4.131	0.667	47.255	0.000
ORANGE	7.100	0.134	4.631	0.667	52.975	0.000
WATRMELN*PINEAPPL	2.400	0.657	0.320	0.667	3.655	0.003
WATRMELN*ORANGE	1.267	0.657	0.169	0.667	1.929	0.078
PINEAPPL*ORANGE	-2.200	0.657	-0.293	0.667	-3.351	0.006

Confidence Interval for Regression Coefficients

Effect	Coefficient	95.0% Confidence Interval		VIF
		Lower	Upper	
WATRMELN	4.600	4.308	4.892	1.500
PINEAPPL	6.333	6.041	6.625	1.500
ORANGE	7.100	6.808	7.392	1.500
WATRMELN*PINEAPPL	2.400	0.969	3.831	1.500
WATRMELN*ORANGE	1.267	-0.164	2.697	1.500
PINEAPPL*ORANGE	-2.200	-3.631	-0.769	1.500

Analysis of Variance

Source	Type III SS	df	Mean Squares	F-ratio	p-value
Regression	9.929	5	1.986	36.852	0.000
Residual	0.647	12	0.054		

Not using a mixture model produces a much larger R (0.999) and an F -value of 2083.371, both of which are inappropriate for these data. Notice that the *Regression Sum-of-Squares* has five degrees of freedom instead of six as in the usual zero-intercept regression model. We have lost one degree of freedom because the predictors sum to 1.

Example 18

Partial Correlations

Partial correlations are easy to compute with GLM. The partial correlation of two variables (a and b) controlling for the effects of a third (c) is the correlation between the residuals of each (a and b) after each has been regressed on the third (c). You can therefore use GLM to compute an entire matrix of partial correlations.

For example, to compute the matrix of partial correlations for $Y1$, $Y2$, $Y3$, $Y4$, and $Y5$, controlling for the effects of X , select $Y1$ through $Y5$ as dependent variables and X as the independent variable.

The input is:

```
GLM
MODEL Y(1 .. 5) = CONSTANT + X
PLENGTH LONG
ESTIMATE
```

Look for the *Residual Correlation Matrix* in the output; it is the matrix of partial correlations among the y 's given x . If you want to compute partial correlations for several x 's, just select them (also) as independent variables.

Computation

Algorithms

Centered sum of squares and cross products are accumulated using provisional algorithms. Linear systems, including those involved in hypothesis testing, are solved by using forward and reverse sweeping (Dempster, 1969). Eigensystems are solved with Householder tridiagonalization and implicit QL iterations. For further information, see Wilkinson and Reinsch (1971) or Chambers (1977).

References

- Chambers, J.M. (1977). *Computational methods for data analysis*. New York: John Wiley & Sons.
- Cochran, W. G. and Cox, G. M. (1957). *Experimental designs*, 2nd ed. New York: John Wiley & Sons.
- Cohen, J. , Cohen, P., West, S.G., and Aiken, L.S. (2002). *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd ed. Hillsdale, N.J.: Lawrence Erlbaum.
- Cornell, J.A. (1985). Mixture experiments. In Kotz, S., and Johnson, N.L. (Eds.), *Encyclopedia of statistical sciences*, vol. 5, 569-579. New York: John Wiley & Sons.
- Dempster, A.P. (1969). *Elements of continuous multivariate analysis*. San Francisco: Addison-Wesley.
- Diamond, W.J. (2001). *Practical experiment designs for engineers and scientists*. 3rd ed. New York: John Wiley & Sons.
- Hocking, R. R. (1985). *The analysis of linear models*. Monterey, Calif.: Brooks/Cole.
- John, P.W.M. (1971). *Statistical design and analysis of experiments*. New York: MacMillan.
- Kutner, M.H, Nachtshiem, C.J., Neter, J., and Li, W. (2004). *Applied linear statistical models*, 5th ed. Irwin: McGraw-Hill.
- * Linn, R. L., Centra, J. A., and Tucker, L. (1975). Between, within, and total group factor analyses of student ratings of instruction. *Multivariate Behavioral Research*, 10, 277-288.
- Milliken, G. A. and Johnson, D. E. (1984). Analysis of messy data, Vol. 1: *Designed Experiments*. New York: Van Nostrand Reinhold Company.
- Morrison, D. F. (2004). *Multivariate statistical methods*, 4th ed. Pacific Grove, CA: Duxbury Press.

- Marquardt, D.W. and Snee, R.D. (1974). Test statistics for mixture models. *Technometrics*, 16, 533-537.
- Searle, S. R. (1971). *Linear models*. New York: John Wiley & Sons.
- Searle, S. R. (1987). *Linear models for unbalanced data*. New York: John Wiley & Sons.
- Wilkinson, J.H. and Reinsch, C. (Eds.). (1971). *Linear Algebra, Vol. 2, Handbook for automatic computation*. New York: Springer-Verlag.
- Winer, B. J., Brown, D. R., and Michels, K.M. (1991). *Statistical principles in experimental design*, 3rd ed. New York: McGraw-Hill.

(* indicates additional references.)

Introduction to Linear Mixed Models

Amit Saxena and Arnab Chakraborty

Linear mixed effects models are useful for analyzing data obtained from designed experiments, for regression analysis and for a host of other situations, where traditional linear models, dealing with only fixed effects, need refinement. These include correlated data, clustered data, dependent data and heteroscedastic data. SYSTAT has provision to analyze various types of linear mixed effects models: variance components models, hierarchical mixed models, and mixed regression. SYSTAT has three different main commands for analyzing linear mixed models: VC for variance components models, MIXED for general linear mixed effects models, and MIX for mixed regression. Their usages are detailed in the subsequent chapters. This chapter is devoted to acquaint the user with the statistical ideas and the theory behind linear mixed models, as well as with the terminology that SYSTAT uses.

Mixed Models and Paired t-test

One way to get started with linear mixed models is by considering paired t-tests as linear mixed model analysis.

Illustrative case: This example is borrowed from a Netmaster online course (see the list of references for the URL). Here we want to compare two methods (High Performance Liquid Chromatography-HPLC and Near Infra Red-NIR) to ascertain the amount of active content in certain tablets. Suppose that we want to test if the two methods yield the same average content. Data have been collected by applying the tests to the same set of 10 tablets (e.g., by breaking each tablet into two halves, and applying one method to each half, assigned at random). The resulting data are shown

in the following table. These data are also available in a SYSTAT file named *TABLET*.

Active content of tablets by two methods:

Tablet	Methods	
	HPLC	NIR
1	10.4	10.1
2	10.6	10.8
3	10.2	10.2
4	10.1	9.9
5	10.3	11.0
6	10.7	10.5
7	10.3	10.2
8	10.9	10.9
9	10.1	10.4
10	9.8	9.9

A standard method to analyze this kind of data is the paired t-test. Let x_i be the measurement by the HPLC method for the i -th tablet, and let y_i be that by the NIR method. Then the paired t-test computes the differences

$$z_i = x_i - y_i,$$

and checks if

$$\frac{Z\sqrt{10}}{\sqrt{\sum_{i=1}^{10} (Z_i - Z)^2/9}}$$

is far from 0 using the t distribution with 9 degrees of freedom, where z is the mean of the Z_i 's.

Let us perform this test using SYSTAT.

The input is:

```
USE TABLET
TESTING
TTEST HPLC NIR
```


The output is:

Paired Samples t-test on HPLC vs NIR with 10 Cases
Alternative = 'not equal'

Mean HPLC	: 10.340
Mean NIR	: 10.390
Mean Difference	: -0.050
95.00% Confidence Interval	: -0.261 to 0.161
Standard Deviation of Difference	: 0.295
t	: -0.535
df	: 9
p-value	: 0.605

This test assumes that z_i 's are independently and identically distributed normal random variables, which is the case if, for example, each (x_i, y_i) pair is independently distributed as $N(\mu_1, \mu_2, \Sigma)$ where Σ is the covariance matrix.

However, if we consider the data set for a moment we can see that Σ cannot be just any covariance matrix. Assuming that both the methods are reasonable, it is highly likely that their measurements for the same tablet will be positively correlated. For instance, if a tablet has a high active content then both the measurements should be high.

Popular as it is, the paired t-test nonetheless fails to take this extra information about the data into account. It collapses the pairs (x_i, y_i) into the differences z_i , and thus fails to utilize the correlation structure of the original data. One way to remedy this loss of information is to assume that each measurement is made up of three components:

- The effect of the actual content of the tablet (which is a random variable depending on the manufacturing process of the tablets.) It is customary to express the effect of the i -th tablet as $\mu + \alpha_i$, where μ is called the mean effect, denoting the average level of active content that an ideal tablet is supposed to contain, while α_i denotes the departure of the i -th tablet from this average.
- The effect of the measurement method (that is where our interest lies.) We shall denote the effect of the j -th method by β_j for $j=1,2$.
- Any random error, which we call ε_{ij}

So we have the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where $i=1, \dots, 10$ and $j=1,2$. Here y_{ij} is the measurement for the i -th tablet obtained by the j -th method. Thus, we have renamed x_i as y_{i1} and y_i as y_{i2}

Readers familiar with the SYSTAT GLM command will quickly recognize this as a linear model. However, there is an important difference between this and the models fit by GLM. In GLM the parameters μ , α_i and β_j are all (unknown) constants. But here the tablet effect α_i 's are random, since the tablets constitute a random sample from the population of all such tablets. The effects μ , β_j are fixed as before. A linear model where some (or all) of the parameters are random is called a *linear mixed model*. Here α_i 's are the random effects, while μ , β_j 's are called fixed effects. We assume that α_i 's and ε_{ij} 's are independent Gaussian (normal) random variables with zero mean. SYSTAT allows various covariance structures for α_i 's and ε_{ij} 's. In this example we shall assume that α_i 's distributed independently as $N(0, \sigma_a^2)$, while ε_{ij} 's have independent $N(0, \sigma_e^2)$ distributions. It is easy to check that the correlation between the two measurements for the same tablet is indeed positive under this model, since

$$\text{Cov}(y_{i1}, y_{i2}) = \text{Var}(\alpha_i) = \sigma_a^2 > 0$$

Let us take a look at how SYSTAT will handles this model. First, SYSTAT demands that the data file be organized in a way that is convenient for this computation as follows:

METHODS	CONTENT	TABLET
HPLC	10.4	1
HPLC	10.6	2
HPLC	10.2	3
HPLC	10.1	4
HPLC	10.3	5
HPLC	10.7	6
HPLC	10.3	7
HPLC	10.9	8
HPLC	10.1	9
HPLC	9.8	10

METHODS	CONTENT	TABLET
NIR	10.1	1
NIR	10.8	2
NIR	10.2	3
NIR	9.9	4
NIR	11.0	5
NIR	10.5	6
NIR	10.2	7
NIR	10.9	8
NIR	10.4	9
NIR	9.9	10

Using the Data=>Reshape=>Stack menu it is easy to convert the data file *TABLET* to this format, rename the columns as METHOD\$, Content (from the default names of Group\$, Variable respectively) and save it as SYSTAT data file *TABLET2*.

The input is:

```
USE TABLET2
VC
CATEGORY TABLET METHOD$
MODEL CONTENT = INTERCEPT + METHOD$
RANDOM TABLET
ESTIMATE
```

Before looking at the output, we point out that our linear mixed model is of a special type called the *variance components model*, and the SYSTAT command for that is VC.

The output is:

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
METHOD\$	1	9	0.287	0.605

Fixed Effects Versus Random Effects

As pointed out in the last section, a mixed linear model is a linear model where some (or all) of the effects are random. These are called the *random effects*, while the others are called *fixed effects*. The randomness in the data is thus split up into two parts: the random effects and the random error. We always assume that the random errors and random effects are independent and are Gaussian with zero mean. The random effects

need not be independent among themselves. The random errors may also be interdependent. Owing to the presence of the random effects the original observations are also correlated. SYSTAT allows different covariance structures for the random effects as well as the random errors, as we shall see later. But first let us see why one would want to consider an effect in a linear model as random.

Illustrative case: Mickey et al. (2004) discuss a data set involving two teaching methods and three teachers. Each teacher uses each teaching method with four different batches of students. The performance of each batch is measured by the average score of the batch in a common examination. The data set is given below:

**Comparing teaching methods
(Scores of Students)**

TEACHER	METHOD 1	METHOD 2
1	67, 73, 59, 84	75, 61, 67, 58
2	92, 84, 94, 83	54, 78, 61, 70
3	74, 72, 76, 64	42, 44, 80, 83

The data are in SYSTAT file *TEACH* (in the format required by SYSTAT in terms of 24 cases (rows) and three columns Score, TEACHER, and METHOD). Let y_{ijk} denote the score of the k -th batch under the i -th teacher using the j -th teaching method. Then y_{ijk} is the resultant of the i -th teacher effect as well as the j -th method effect. In fact, we can also take into account the batch effect, but for this study we shall assume that the batches are all more or less identical. Also, we shall ignore any interaction between teacher and method. (One can actually include the interaction to satisfy oneself that the interaction is insignificant.) So we have the linear model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Here μ is the mean effect, α_i is the i -th teacher effect, and β_j is the effect of the j -th method. The ε 's, as usual, denote the random errors.

Now let us pause for a moment and wonder why one would really collect and analyze a data set of this kind. In other words, what type of inference do we want to make? There are two possible answers to this.

First, we may be interested in knowing how these three teachers perform using the two methods. This question is of interest to, for instance, the head of a school, when she wants to decide which method to adopt. Here she has a *specific* set of teachers in mind.

Second, an educator may want to compare the two teaching methods irrespective of the teachers. He does not have any specific set of teachers in mind. He is comparing the performance of method 1 as applied by some randomly selected teacher, with the performance of method 2 applied by another (possibly different) randomly selected teacher.

In the first case all the effects are fixed. In the second case, the teacher effects α_i 's are random. Let us analyze the data set under both the models to see how the inference differs. First, the fixed effects model.

The input is:

```
USE TEACH
VC
MODEL SCORE = INTERCEPT + TEACHER + METHOD
ESTIMATE
```

Notice how we have used the VC command even though we are fitting a fixed effect model. This is because fixed effects models are special cases of mixed effects models where there are no random effects. We could also have used the GLM commands.

The input is:

```
USE TEACH
GLM
MODEL SCORE = INTERCEPT + TEACHER + METHOD
ESTIMATE
```

Either method produces the same information. We show a relevant part of the output from the VC command.

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	90.375	10.238	21	8.828	0.000
TEACHER	-0.563	3.135	21	-0.179	0.859
METHOD	-12.417	5.119	21	-2.426	0.024

Type III Tests for Fixed Effects

Source	Numerator df	Denominator df	F-ratio	p-value
TEACHER	1	21.000	0.032	0.859
METHOD	1	21.000	5.884	0.024

Next we apply the mixed effects model where the teacher effect is random.

The input is:

```
USE TEACH
VC
MODEL SCORE = INTERCEPT + METHOD
RANDOM TEACHER
ESTIMATE
```

A relevant snippet from the output is shown below.

Estimates of Covariance Components

Random Effect	Description	Estimate
TEACHER	Variance	0.010
	Parameter	
Error variance	Variance	150.299
	Parameter	

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	89.251	7.916	21	11.275	0.000
METHOD	-12.417	5.005	21	-2.481	0.022

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
METHOD	1	21	6.155	0.022

Notice that the result of the analysis is now different: the p -value has gone down. This means that the methods appear more significantly different when used over a population of teachers, than when used for just a specific set of teachers. It could have been the other way around also. Then the interpretation would be as follows.

- The significant difference in the fixed effects model implies that if the same teacher uses both the methods then the results are different.
- The lack of significance in the mixed effects model means that a random teacher using one method has more or less the same performance as a (possibly different) random teacher using the other method. This is the case if, for instance, there is a lot of variability among the teachers, and the difference between the teaching methods is swamped out by it. A bad teacher with a good method may not perform much differently from a good teacher with a bad method.

Why Use Random Effects?

A linear model, just like any other statistical model, tries to capture the essence of the *process generating the data*, rather than that of the data itself. We want our inference to hold not only for the given data set but also for future replications of the same experiment. So the choice of the model is dictated by what type of replications we have in mind. Depending on this there are different reasons behind treating an effect as random in a model. Here we outline three such common situations.

- If we plan to use the same levels of the some effect in all fresh replications, then we may treat the effect as fixed. However, if we plan to use fresh levels of some effect, then we should make the effect random. Inference based on random effects models are valid for a population of all possible levels of the random effects. The teaching method example furnished one illustration. In such situations, the random coefficients are all independently and identically distributed, as they represent randomly selected levels of the effect. So the resulting model is a variance components model. The next chapter has many more examples of such models in use for real life data sets.
- In some cases, an effect may be considered random even if we plan to use the same levels for all future replications. Consider, for instance, a designed experiment where 3 operators in a factory are operating 2 machines, the response being a score that combines the quality and quantity of the output produced in a given amount of time. A suitable model for this situation may be:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the score for the k -th run of the i -th machine operated by the j -th operator. If the factory has only these three operators to operate the machines, then the factory authorities would have to always choose the same three operators in all future replications of the experiment. However, the same operator may behave slightly differently from one replication of the experiment to the next depending on unpredictable factors like his mood. In this case, we would be justified to consider the operator effect as random. However, since the mood variability of the different operators may be different, so here the random coefficients β_j 's need not be identically distributed. In fact, they may also be correlated, because the moods of the all the operators may be affected by some common random condition prevailing during a replication of the experiment (e.g., weather during the experiment) that is difficult to

control. Indeed, McLean et al. (1991) also suggests a model where the operator effect is fixed, but the interaction effect (γ_{ij}) is random. Such a model would be appropriate if we consider the main effect as a measure of the proficiency of the operator, which is not likely to change between replications. However, the mood fluctuations may affect how an operator operates a given machine. Such models where the random effect coefficients may not be independently and identically distributed are more general than simple variance components models. The MIXED command is designed to tackle these cases. Real life examples are furnished in the chapter Linear Mixed Models and Hierarchical Linear Mixed Models.

- A third situation that leads to random effects is where the model is developed in a multi-level fashion. Consider a situation where we want to linearly regress a response variable y on a predictor variable x . However, we believe that the regression slope is a random effect that depend on the values of a categorical variable z . Then we have a two-level model. In the first level we model y in terms of x :

$$y_{ijk} = \alpha + \beta_j x_{ijk} + \varepsilon_{ijk}$$

Here j denotes the levels of the categorical variable z . In the second level we model the (random) regression slope in terms of z :

$$\beta_j = a + b_j$$

Here b_j 's are random effect coefficients. Putting the second level equation in the first we get the composite model:

$$y_{ijk} = \alpha + (a + b_j)x_{ijk} + \varepsilon_{ijk}$$

This means that here x is present in the fixed part ax_{ijk} as well as in the random part $b_j x_{ijk}$ effect. If the deeper levels in a multi-level model have their own random errors, then they lead to random effects in the composite model. The SYSTAT specification for the above example is:

```
MODEL Y = INTERCEPT + X
```

```
RANDOM X / GROUP = Z
```

Such models often have the same effect in both the MODEL and RANDOM lines, as x is in this example. Milliken and Johnson (1992) have more examples of a like nature. The chapter Linear Mixed Models in this manual illustrates one such model in action.

Some Linear Model Terminology

Now we shall discuss some important issues about linear models, and how SYSTAT handles them in the context of mixed effects models. Users familiar with the GLM command of SYSTAT may just like to skim through this section.

String and Numeric Variables

A variable in SYSTAT can be either a *string* or *numeric*. The names of string variables end with a \$ sign. The values taken by a string variable can be numeric or alphanumeric, but the numbers will only be interpreted as names or symbols and not usable for calculations. Thus string variables are used as categorical variables. Numeric variables take numbers as their values. However, you can ask SYSTAT to treat a numeric variable as categorical by using the CATEGORY command. For instance, the command

CATEGORY X Y, treats x and y as categorical numeric variables. The command CATEGORY

in a line by itself restores all numeric variables to their default continuous status. Thus string variables represent categorical variables can be represented by numeric variables also. Numeric variables can also represent discrete or continuous variables. SYSTAT does not differentiate between discrete and continuous variables.

Estimability

Consider the model,

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i=1,2,3$ and $j=1,2$. It is a well known fact from linear models theory that the parameters μ and α_i 's are not estimable from the data unless we impose further restrictions on the parameters. One possible restriction is to assume that sum of the α_i 's equals 0. Then μ measures the overall average, while α_i 's measure the departure of the i -th group average from the overall average. Another popular restriction is $\alpha_3 = 0$. In this case, μ measures the average of the third group, while α_1 and α_2 measure the departure of the first and second group averages from that of the third group.

SYSTAT calls the first restriction as *effects* encoding and the second restriction as dummy encoding. SYSTAT uses effects encoding for the fixed effects. However, as we shall see later, the random effects coefficients do not suffer from any estimability problem. So we leave them unconstrained. This is called *means* encoding.

Data Layout: Multiway or Nested

Linear models seek to explain the variation present in the data in terms of various effects. The part that cannot be explained thus is ascribed to random error. For instance, variation in the yields of a crop may be partly explained by the effects of fertilizers and soil type. When more than effect is present their combination determines the layout of the data. There are two basic layouts: *multiway* and *nested*.

In a multiway layout all the values of each categorical variable have the same meaning irrespective of the values of the other categorical variables. The following is an example of a 2-way layout, i.e., multiway layout with just two effects.

Illustrative case: Consider a study to investigate whether the IQ level of a person depends on the person's gender and lefthandedness. A typical data set will look like the following where y_i is the IQ of the i -th person.

	Male	Female
Lefthanded	y_1, y_2	y_3, y_4
Righthanded	y_5, y_6	y_7, y_8

Here left-handedness means the same thing for both males and females. So in this is a 2-way layout.

Data sets having 2-way (or even multiway) layouts are often presented as tables like the above, where rows (and, if necessary, subrows) are devoted to some effect(s), while columns (and, if necessary, subcolumns) are devoted to other effect(s). Such a table facilitates human reading, but SYSTAT always expects its input in a table where each experimental unit has its own row, and each variable has its own column. Thus, the above data set must be presented in the following format to SYSTAT.

Gender\$	Handedness\$	IQ
Male	Left	Y ₁
Male	Left	Y ₂
Male	Right	Y ₃
Male	Right	Y ₄
Female	Left	Y ₅
Female	Left	Y ₆
Female	Right	Y ₇
Female	Right	Y ₈

Multiway layouts can be of two general types: *additive* and *non-additive*. These are also called models without interaction and with interaction, respectively. In SYSTAT interaction is called *crossing*.

Illustrative case: Consider an agricultural experiment, where we want to relate the yields of crops to the soil type and the type of fertilizer used. Here are two possible hypothetical datasets. The data are in SYSTAT files *AGR1* and *AGR2* respectively.

A data set on agricultural yield:

Fertilizer	Soil 1	Soil 2	Soil 3
1	10, 12	34, 30	20, 23
2	5, 4	29, 28	14, 16

Another data set on agricultural yield:

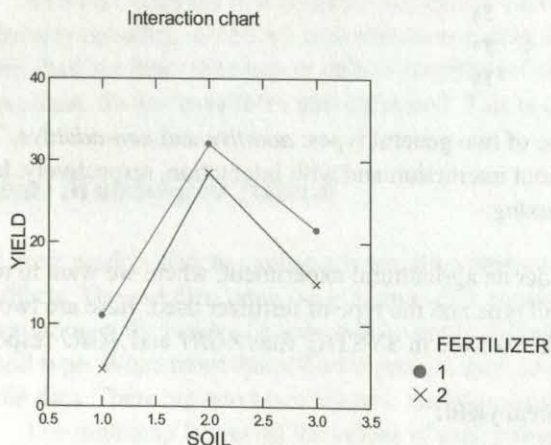
Fertilizer	Soil 1	Soil 2	Soil 3
1	10, 12	34, 30	20, 23
2	30, 31	21, 16	19, 25

It is always a good idea to look at any data graphically before performing formal statistical analyses. So let us plot the two data sets as follows. We have chosen to plot the average yield against Soil and have used different colors for different types of Fertilizers. The data file *AGR1* and *AGR2* have been recast in the format required by SYSTAT and the columns have been named YIELD, FERTILIZER, and SOIL.

The input is:

```
USE AGR1
SSAVE MEANS_Y
BY SOIL FERTILIZER
CBSTAT YIELD / MEAN
USE MEANS_Y
DOT YIELD*SOIL/OVERLAY GROUP=FERTILIZER LINE,
      TITLE="Interaction chart"
```

The output is:



The plot shows that the lines for different fertilizers are more or less parallel. In other words, the general form of relation between yield and soil types is the same for all fertilizers. However, each fertilizer has an *additive* effect that shifts the yield-vs-soil curve up and down. This is a case where an *additive* model is a good choice. Here we

can easily compare between the fertilizers: the first is clearly better than the second. In SYSTAT this model will be written as:

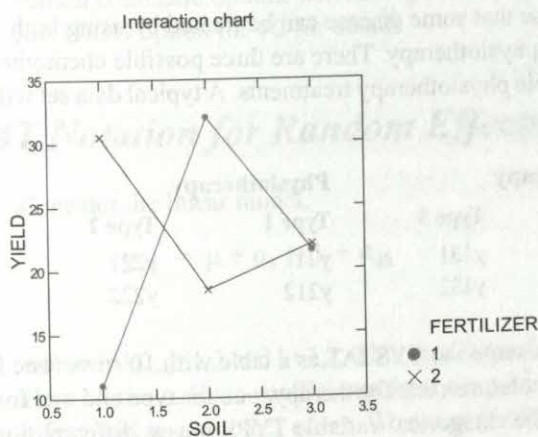
```
USE AGR1
MIXED
  CATEGORY FERTILIZER SOIL
  MODEL YIELD = INTERCEPT + FERTILIZER + SOIL
ESTIMATE
```

Note our use of the MIXED command. Actually, we could also have used the simpler VC command. This input corresponds to additive model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where α_i 's are the fertilizer effects, β_j 's are the soil effects, and μ is the overall mean effect.

Next let us take a look at the interaction chart for the second data set.



Here the situation is quite different. In this case we cannot make a clear statement about which fertilizer is better, the first fertilizer fares well for some soil types, while the second fertilizer is better for the other soil types. This is manifested through the non-parallel nature of the two lines in the interaction chart. We say that there is *interaction* between soil type and fertilizer. This calls for a model with *interaction*:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

which is the same as the last model, except for the interaction effects, γ_{ij} 's. Indeed, this model subsumes the additive model as a special case with $\gamma_{ij}=0$.

The input is:

```
USE AGR2
MIXED
  CATEGORY FERTILIZER SOIL
  MODEL YIELD = INTERCEPT + FERTILIZER + SOIL + FERTILIZER*SOIL
ESTIMATE
```

Nested Layout

In some data sets the values of one effect A assume different meanings for different values of another effect B. Then we say that A is nested inside B.

Illustrative case: Suppose that some disease can be treated by using both chemotherapy as well as physiotherapy. There are three possible chemotherapy treatments and two possible physiotherapy treatments. A typical data set will then look like

	Chemotherapy		Physiotherapy	
Type 1	Type 2	Type 3	Type 1	Type 2
y111	y121	y131	y211	y221
y112	y122	y132	y212	y222

This data set should be presented to SYSTAT as a table with 10 rows (one for each experimental unit) and 3 columns (one for therapy, one for type and one for the response variable.) Here the categorical variable TYPE means different things for Chemotherapy and Physiotherapy. We say that the TYPE effect is nested inside the THERAPY effect.

Statistics textbooks often use the following model for this situation:

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

where α_i denotes the effect of the i -th therapy and $\beta_{j(i)}$'s stand for the nested effect of the j -th type inside the i -th therapy. This model is written in SYSTAT syntax as follows:

```
MODEL Y = INTERCEPT + THERAPY + TYPE(THERAPY)
```

Balanced and Unbalanced Data

A typical linear model data set consists of numbers that are classified into cells. In an agricultural study, for instance, the data may be yields of different crops, where the cells defined by the combinations of crop type, fertilizer type. If each cell in a cross model contains an equal number of observations then we call the data set as *balanced*. Such datasets are usually easier to analyze and interpret. SYSTAT allows the user to deal with both balanced and unbalanced data. The user does not need to explicitly specify whether the data set is balanced or not, since SYSTAT can figure that out from the data itself. In fact, one beauty of mixed models is that they provide a unified framework for dealing with both balanced and unbalanced data. However, there are certain command options that take special effect for unbalanced data sets. See the METHOD option for VC for details.

SYSTAT Notation for Random Effects

Consider the linear model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where $i=1,2$, $j=1,2$ and $k=1,2$. This model has 5 coefficients: 1 μ , 2 α_i 's and 2 β_j 's. Here all the α_i 's are coefficients for the same effect, all the β_j 's belong to another effect, and μ is another effect. We can write this in matrix notation as:

$$Y = X\beta + \varepsilon$$

where

$$X = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \end{bmatrix} \quad y = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \end{bmatrix}$$

In a mixed effects model each effect is either fixed or random. The coefficients corresponding to a fixed effect are treated as parameters. The coefficients for a random effect are assumed to be random variables that follow Gaussian distribution with mean zero and some user-specified covariance structure. For instance, we may want to treat the β_j 's as random, keeping μ and α_i 's fixed. Then it is customary to write the fixed part and the random part separately in matrix notation, as discussed below. This motivates the following general definition of mixed effects models.

A linear mixed model is a linear model of the form:

$$Y = X\beta + Z\gamma + \varepsilon$$

where Y is the data vector, X and Z are known matrices (either design matrices or covariate matrices), β is the vector of fixed effects, γ is the vector of random effects, and ε is the random error vector. Here Y is a random vector, whose randomness comes partly from the random vector γ and partly from ε . We assume that

$$\begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} \right)$$

In particular, γ and ε are independent. We are denoting $\text{Var}(\gamma)$ by G and $\text{Var}(\varepsilon)$ by R . Depending on the choices of X , Z , G , and R we have different types of linear mixed models. We shall take a brief tour through these shortly. To declare an effect as random we use the **RANDOM** command:

```
MODEL Y = CONSTANT + A
RANDOM B
```

It is possible to have multiple effects in the same **RANDOM** line and/or multiple **RANDOM** lines in the same model. But before learning about them we need to understand how SYSTAT specifies the structures of G and R .

Covariance Structures

The model and covariance matrices for the random effects and random errors determine the covariance matrix of the data set as follows:

$$\text{Var}(Y) = Z'GZ + R$$

SYSTAT gives the user a choice from certain standard types of covariance structures for G and R . There are four choices for the structure of G and two choices for that of R . These are listed below. To illustrate these we use the following hypothetical data set as our running example. These data are in SYSTAT file *COVSTRUCT*.

To specify the structure of the covariance matrices A and B we use the **STRUCTURE** option for the **RANDOM** and **REPEATED** commands, respectively, like this:

```
USE COVSTRUCT
MIXED
  CATEGORY P Q
  MODEL Y = INTERCEPT + P
  RANDOM Q / STRUCTURE = CS
  REPEATED / STRUCTURE = VC
ESTIMATE
```


The values CS and VC specify the structures of the covariance matrices. These and the other possible structures are described below.

- **Variance Components (VC).** Here the covariance matrix is a diagonal matrix with equal diagonal entries, i.e., a matrix of the form $\sigma^2 I$, where I is the identity matrix of appropriate size. The 4 by 4 case is shown below,

$$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Notice that this structure has exactly one parameter irrespective of the size of the matrix. Change the RANDOM line in the running example to

```
RANDOM Q / STRUCTURE=VC
```

to produce the output

Estimates of Covariance Components

Random Effect	Description	Estimate
Q	Variance	0.001
	Parameter	
Error variance	Variance	1.948
	Parameter	

VC is the default value for the STRUCTURE option for both the RANDOM and the REPEATED command. Thus the above SYSTAT line could be abbreviated to just RANDOM Q.

- **Compound Symmetry (CS).** Here the diagonal entries of the covariance matrix of the observations are all same, and so are the off-diagonal entries. If we write, for instance,

```
RANDOM Q / STRUCTURE=CS
```

we get the following structure for the covariance matrix:

$$\begin{bmatrix} \sigma^2 & T & T & T \\ T & \sigma^2 & T & T \\ T & T & \sigma^2 & T \\ T & T & T & \sigma^2 \end{bmatrix}$$

This has two parameters irrespective of the size. In the output of the MIXED command the T parameter is called Compound Symmetry.

Estimates of Covariance Components

Random Effect	Description	Estimate
Q	Variance	0.000
	Parameter	
	Compound	0.000
	Symmetry	
Error variance	Variance	1.948
	Parameter	

Notice that we do not have any Compound Symmetry row for the error variance, because it is still using default VC covariance structure. We can of course use CS structure there as well:

```
RANDOM Q / STRUCTURE = CS
REPEATED / STRUCTURE = CS
```

Random Effect	Description	Estimate
Q	Variance	0.780
	Parameter	
	Compound	0.560
	Symmetry	
Error variance	Variance	1.701
	Parameter	
	Error	-0.067
	Correlation (CS)	

Observe that the t for the error covariance matrix is not printed directly. Rather it is divided by the error variance, to produce the error correlation.

- **Diagonal (DIAG).** Here all the observations are uncorrelated but may have different variances. Thus, for this option the number of covariance parameters equals the size of the matrix.

$$\begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

If we use

RANDOM Q / STRUCTURE=DIAGONAL

we get the output

Random Effect	Description	Estimate
Q	Variance 1	0.000
	Variance 2	1.239
	Variance 3	0.000
	Variance 4	0.000
Error variance	Variance	1.683
	Parameter	

- **Unstructured (UN).** This case, as the name suggests, does not put any restriction on the covariance matrix (except, of course, that the matrix should be positive definite.)

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

If we use

RANDOM Q / STRUCTURE=UN

We get the output:

Estimates of Covariance Components

Random Effect	Description	Estimate
Q	Variance 1	0.984
	Covariance(2, 1)	0.307
	Variance 2	1.456
	Covariance(3, 1)	0.800
	Covariance(3, 2)	0.638
	Variance 3	0.768
	Covariance(4, 1)	0.868
	Covariance(4, 2)	0.521
	Covariance(4, 3)	0.778
	Variance 4	0.814
Error variance	Variance Parameter	1.653

Notice that only the lower triangular half of the symmetric covariance matrix is reported.

Illustrative case: Consider the SYSTAT commands:

```
MIXED
CATEGORY A B
MODEL Y = CONSTANT
RANDOM A B A*B / STRUCTURE = VC
ESTIMATE
```

Here the random errors are uncorrelated, and so are all the random effects. The random errors have a common variance. The random effects of the same type have a common variance, which may be different from the common variance of the random effects of a different type. For instance, consider the model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},$$

where α_i , β_j 's and γ_{ij} 's are all random effects. If we use the VC structure, then the ε_{ijk} 's are uncorrelated with a common variance σ_e^2 ; α_i 's are uncorrelated with common variance σ_a^2 ; β_j 's are uncorrelated with common variance σ_b^2 ; and γ_{ij} 's are uncorrelated with common variance σ_g^2 . Thus, under this model,

$$\text{Var}(y_{ijk}) = \sigma_a^2 + \sigma_b^2 + \sigma_g^2 + \sigma_e^2$$

More on RANDOM

We can write in the RANDOM line any effect that can be written in the MODEL line. Thus, all the following are valid:

```
RANDOM INTERCEPT
RANDOM P
RANDOM P*Q
RANDOM P(Q)
```

It is possible to list multiple effects in the same RANDOM line. Any STRUCTURE option in such a line applies to the multiple effects jointly. For instance

```
RANDOM P Q / STRUCTURE=CS
```

clubs the 2 coefficients for P and 4 coefficients for Q in a single coefficient vector of length 6, and postulates a 6x6 covariance matrix of the compound symmetric type. So here we have just two covariance parameters for all the 6 coefficients. This is easy to see from the following output. Estimates are all close to zero.

Estimates of Covariance Components

Random Effect	Description	Estimate
P + Q	Variance	0.000
	Parameter	
	Compound	0.000
	Symmetry	
Error variance	Variance	1.948
	Parameter	

But if we use,

```
RANDOM P / STRUCTURE = CS
RANDOM Q / STRUCTURE = CS
```

then the covariance matrix of the 6 coefficients is block diagonal, one block for P, the other for Q. Each block is of the compound symmetric type, with its own parameters. Thus, here we have four covariance parameters in all. The relevant part of the output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
P	Variance	1.733
	Parameter	
	Compound	0.001
	Symmetry	
Q	Variance	0.000
	Parameter	
	Compound	0.000
	Symmetry	
Error variance	Variance	1.948
	Parameter	

Finally, SYSTAT allows block diagonal matrices where all the blocks are identical. Examples are

```
RANDOM P / GROUP= Q STRUCTURE = CS
```

Here we are nesting P inside Q. So we have $2 \times 4 = 8$ coefficients. We are grouping the coefficients into four groups by the value of Q. The coefficients in different groups are assumed independent in this model. Also each group has the same covariance matrix. Thus the covariance matrix of all the 8 coefficients is an 8x8 block diagonal matrix consisting of four equal blocks of size 2, and each block has a compound symmetric structure.

This is different from the model:

```
STRUCTURE=CS
or
RANDOM P*Q / STRUCTURE = CS
```

both of which have 8 coefficients, but the covariance matrices are not block diagonal, rather they are 8x8 compound symmetric matrices.

It is also possible to have the GROUP option in the REPEATED command. For instance,

```
REPEATED / GROUP = P
```

Using Covariates: Regression

All our examples so far deal with categorical variables. However, SYSTAT can also handle the case where one or more explanatory variables are real. These correspond to regression situations. If only X matrix has real variables, but Z is a design matrix, then we have the usual regression set up. If, on the other hand, there are real variables in Z, we have a mixed regression situation. We illustrate these below.

Illustrative case: First let us perform a simple linear regression analysis using MIXED. It is certainly an overkill to use MIXED for a simple task like this, but it makes a good introductory example. Consider the following hypothetical data set in the SYSTAT file *HW*.

Height-Weight data.

GENDER	HEIGHT (Inches)	WEIGHT (Kgs.)
MALE	6.2	76
MALE	5.8	68
MALE	5.0	60
MALE	5.6	58
MALE	5.8	69
FEMALE	5.3	70
FEMALE	5.2	65
FEMALE	5.5	69
FEMALE	5.7	59
FEMALE	5.2	62

Initially we shall ignore GENDER, and try to fit a linear regression of WEIGHT on HEIGHT.

The input is:

```
USE HW
MIXED
  MODEL WEIGHT = INTERCEPT + HEIGHT
ESTIMATE
```

The output is:

Analysis of Variance

Source	Type III SS	Numerator df	Denominator df	Mean Squares	F-ratio	p-value
HEIGHT	86.718	1	8.000	86.718	3.217	0.111
ERROR	215.682		8	26.960		

Fit Statistics

Final L-L	: -25.763
-2L-L	: 51.527
AIC	: 53.527
AIC (Corrected)	: 54.194
BIC	: 53.606

Estimates of Variance Components

Source	Variance Components	Standard Error	Z	p-value	95.00% Confidence Interval Lower	Upper
Error	26.960	13.480	2.000	0.046	0.540	53.381

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	18.213	26.473	8	0.688	0.511
HEIGHT	8.569	4.778	8	1.793	0.111

However, if we want to fit different lines for different genders, then we should change the MODEL line to,

$$\text{MODEL WEIGHT} = \text{GENDER\$} + \text{GENDER\$*HEIGHT}$$

A relevant snippet from the output is shown below.

Analysis of Variance

Source	Type III SS	Numerator df	Denominator df	Mean Squares
GENDER\$*HEIGHT	90.734	2	7.000	45.367
ERROR	211.666		7	30.238

Analysis of Variance (contd...)

Source	F-ratio	p-value
GENDER\$*HEIGHT	1.500	0.287
ERROR		

Fit Statistics

Final L-L	: -25.146
-2L-L	: 50.292
AIC	: 52.292
AIC (Corrected)	: 53.092
BIC	: 52.238

Estimates of Variance Components

Source	Variance Components	Standard Error	Z	p-value	95.00% Confidence Interval	
					Lower	Upper
Error	30.238	16.163	1.871	0.061	-1.441	61.917

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t
Intercept		12.858	31.654	7	0.406
GENDER\$*HEIGHT	FEMALE*HEIGHT	9.670	5.894	7	1.641
	MALE*HEIGHT	9.412	5.563	7	1.692

Estimates of Fixed Effects (contd...)

Effect	Level	p-value
Intercept		0.697
GENDER\$*HEIGHT	FEMALE*HEIGHT	0.145
	MALE*HEIGHT	0.135

Similarly, the model,

$$\text{MODEL WEIGHT} = \text{INTERCEPT} + \text{GENDER\$*HEIGHT}$$

fits two regression lines with a common intercept, while the model,

$$\text{MODEL WEIGHT} = \text{GENDER\$} + \text{HEIGHT}$$

fits two regression lines with a common slope.

Next let us consider a model with common slope, but where the intercept terms are random. The SYSTAT command lines are

```
USE HW
MIXED
  MODEL WEIGHT = INTERCEPT + HEIGHT
  RANDOM GENDER$
ESTIMATE
```

The output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
GENDER\$	Variance	0.002
	Parameter	
Error variance	Variance	26.960
	Parameter	

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	18.211	26.474	1	0.688	0.616
HEIGHT	8.569	4.778	7	1.793	0.116

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
GENDER\$	FEMALE	0.000	0.048	7	0.006	0.995
	MALE	0.000	0.048	7	-0.006	0.995

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
HEIGHT	1	7	3.217	0.116

Estimation and Prediction

The covariances and the fixed effect coefficients are the parameters to be estimated in a linear mixed model. Besides these estimations, SYSTAT can also predict the random effect coefficients. SYSTAT computes Best Linear Unbiased Estimators (BLUE) for the fixed effects, and Best Linear Unbiased Predictors (BLUP) for random effects. Details are given below. SYSTAT offers a number of methods to compute the covariance matrices: three types of analysis of variance (ANOVA) methods, Minimum Variance Quadratic Unbiased Estimation (MIVQUE0), maximum likelihood method and Restricted or Residual Maximum Likelihood (REML) method. Among these, the ANOVA and MIVQUE0 methods are applicable only to the models analyzed by the VC command.

Estimating the Fixed Effects

SYSTAT's estimation of the fixed effect parameters produce Best Linear Unbiased Estimators (BLUE), which has a number of desirable properties. By "Linear" we mean that the estimator scales with the input. For instance, if in a financial data set the unit is changed from Dollar to Euro, then the estimator will also be scaled by the Dollar-Euro exchange rate. The estimator is also unbiased and among all linear unbiased estimators it is the best in the sense that it has the minimum possible variance.

The BLUEs and BLUPs are obtained by solving Henderson's linear Mixed Model Equations (MME). For justification and algorithmic details please refer to the Computational Details section at the end of this chapter.

Some salient aspects of BLUEs:

- They are a form of Weighted Least Squares (WLS) solution. These are better than Ordinary Least Square (OLS) estimators if the data are correlated.
- The weights are proportional to the precision of the estimates. Hence sometimes these estimators are called precision-weighted estimators.

Predicting the Random Effects

Next we discuss prediction of random effects. This prediction is done using the conditional expectation of the random effects given the observations (Y). These predictions are linear functions of Y and have minimum mean-square error in the class of all linear, unbiased predictors. Hence they are called Best Linear Unbiased Predictors (BLUP). The BLUPs are obtained from the solutions of Henderson's mixed model equations (Henderson, Kempthorne, Searle, and von Krosig, 1959).

Some important properties of these BLUPs are as follows:

- Under normal priors BLUPs are Empirical Bayes estimates.
- They are a form of *shrinkage estimates*. Such a predictor has lower variance than the estimator obtained by treating the effect as fixed.

We shall illustrate these in the Further Insights section at the end of this chapter.

Standard Errors

No point estimate should ever be quoted without mentioning its standard error. This gives an idea how different the estimate could be if we replicate the experiment. In the presence of random effects *replicating an experiment* can have more than one interpretation, because we have two sources of randomness: the random errors and the random effects. Accordingly we can have different types of replication:

- We freshly randomize all the random effects (*Broad Inference Space*)
- We hold all the random effects fixed at their levels in the original data set (*Narrow Inference Space*)
- We randomize some of the random effects, and hold rest fixed at the present levels (*Intermediate Inference Space*)

If we do not specify otherwise SYSTAT reports only the broad inference space standard errors of the BLUEs and BLUPs. It is possible to make SYSTAT compute

standard errors using narrow or intermediate inference spaces also. We shall discuss this in the hypotheses testing section in this chapter.

Estimating Covariance Matrices

As we have already touched upon, there is more than one way to estimate the covariance matrices using SYSTAT:

- Analysis of variance (ANOVA) method: this has three flavors: TYPE1, TYPE2 and TYPE3. All these are applicable only for VC models.
- Minimum Variance Quadratic Unbiased Estimation (MIVQUE0). This is also applicable only for VC models.
- Maximum Likelihood (ML) method
- Restricted or Residual Maximum Likelihood (REML) method

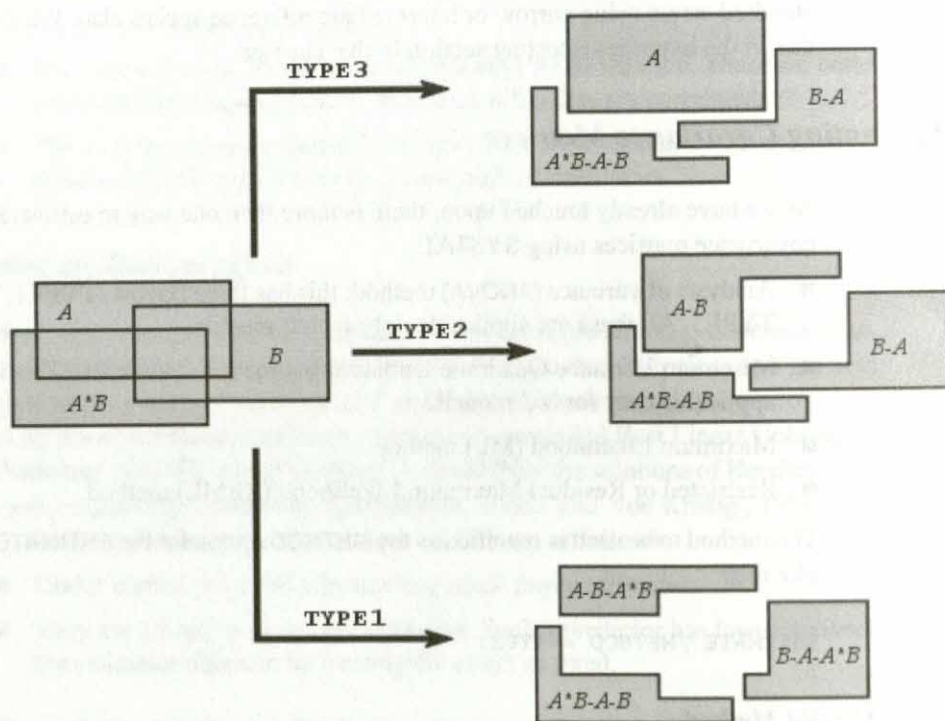
The method to be used is specified as the METHOD option for the ESTIMATE command, like this:

```
ESTIMATE /METHOD = TYPE3
```

ANOVA Method

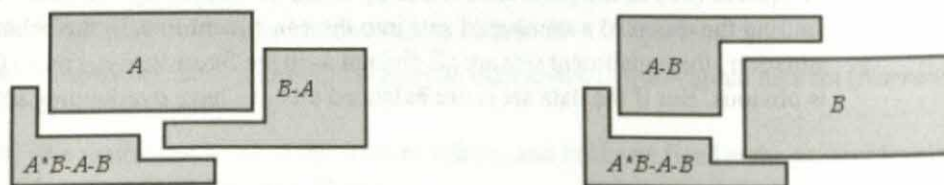
This is a special case of method of moments estimation, where we equate the mean sum of squares (MS) to their expectations, and solve the resulting system of equations. This method has three different versions for unbalanced data, i.e., where the numbers of cases in the different cells are not the same. The three versions are commonly called Type1, Type2 and Type3. We explain these below.

In a variance components model we try to break the original data vector into different parts corresponding to the different effects and random error. Ideally the sum of squares (SS) of the parts should add up to the SS of the original data. This is like spitting the union of a number of sets into the constituent sets. In the balanced data situation, the constituent sets are all disjoint as in the figure below, and so the splitting is obvious. But if the data set is not balanced then we have overlapping sets.



The 3 Types of Sum of Squares

Note that in Type 1, the SS for each effect is computed after taking out the contributions of all other effects. In Type 3, we proceed sequentially: first A, then B sans the contribution of A, then A*B sans the contribution of A and B. Owing to this sequential nature, Type 3 SS's depend on the order in which the effects are listed.



Type 2 is suggested as a compromise between Types 1 and 3 to achieve symmetry. Here, for each effect we take out the contributions of the effects of same or lower

orders only. For instance, the Type 2 SS for A in our example is computed as A minus B. The SS for B is similarly, B minus A. The interaction, being a higher order term, does not enter into the picture so far. To compute the SS for A*B we need to take out the effects of both A and B. For an example of the three types of SS in action, please see the Variance Components chapter in this manual.

Here is a command line that requests the Type 1 method.

```
ESTIMATE /METHOD = TYPE1
```

You may replace the TYPE1 keyword by TYPE2 or TYPE3 to specify the ANOVA estimation method of your choice. The default is TYPE3 for models analyzed by VC.

There is no consensus among statisticians as to which method is the best. While the scale tilts more toward type 3, the other two methods have their share of ardent supporters as well (e.g., Milliken and Johnson (1992).) The controversy stems from the fact that the different types of SS's actually test different hypotheses. The hypotheses tested by Types 1 and 2 involve the unequal cell frequencies or unbalanced cases: a fact that is vehemently disparaged by many on the ground that hypotheses should be statements involving only the model and not the data. Type 3 hypotheses are free of this blemish. However, supporters of the first two types argue that the dependence of a hypothesis on the cell frequencies may be justified if the cell frequencies reflect the underlying population sizes. Also, the hypotheses tested by the Type 3 method are sometimes less intuitive to interpret.

Some salient aspects of the ANOVA methods of estimating variance components are:

- They not dependent on the distributional assumptions on the effects. They use information about only the first two moments.
- These are non-iterative methods , and usually require less computation than iterative methods.
- ANOVA estimates of variance components can be negative.
- ANOVA methods in SYSTAT are applicable only to variance components models analyzed by VC.

MIVQUE0

MIVQUE0 was originally proposed by Rao (1971) to estimate the variances in a variance components model without using any normality assumptions. This is a special

case of Minimum Norm Quadratic Unbiased Estimation (MINQUE) procedure. First, let us write the variances as a column vector,

$$s = (\sigma_1^2, \dots, \sigma_k^2, \sigma_e^2)'$$

Our aim is to estimate some linear combination $p's$, where p is a known vector. For instance, if we want to estimate σ_e^2 we shall take $p = (0, \dots, 0, 1)'$.

In the general MINQUE method, we start with a guess s_0 for s . Let V_0 be the covariance matrix of Y using the guess s_0 in place of s . Then we look for a quadratic function $Y'QY$ of the data Y to minimize:

$$\text{trace}[(QV_0)^2]$$

with respect to Q such that

- $Q=Q'$
- $QX=0$ (for translation invariance)
- $\pi = \text{trace}(QZiZi')$ (for unbiasedness)

In MIVQUE0 we take $s_0 = (0, \dots, 0, 1)$

The SYSTAT syntax for MIVQUE0 is:

```
ESTIMATE /METHOD = MIVQUE0
```

Some salient aspects of the MIVQUE0 method of estimating variance components are:

- MIVQUE0 is applicable only for variance components models.
- MIVQUE0 is a non-iterative method.
- MIVQUE0 may produce negative variance component estimates. This is a general problem with method of moments estimators. They may produce estimates falling outside the parameter space..

Maximum Likelihood (ML)

To request ML estimation in SYSTAT, the input is:

```
ESTIMATE /METHOD = ML
```


Some salient aspects of the ML method of estimating covariance parameters are:

- These estimators are consistent, asymptotically efficient, and asymptotically normal under quite general assumptions
- Asymptotic covariance matrix of the estimators is produced as a by product. It is the inverse of the information matrix.
- SYSTAT implements ML estimation using an iterative algorithm.
- ML estimation in SYSTAT depends on normality assumptions.
- ML estimators are usually biased.

Restricted or Residual Maximum Likelihood (REML)

While likelihood based methods have many charming properties (e.g., asymptotic normality, achieving lowest possible variance asymptotically), ML estimators suffer from one drawback: they are biased estimators in general. For mixed effects models the origin of this bias may be explained as follows. The ML method first estimates the fixed effects, and then estimates the covariance parameters by treating the residuals as fresh data. However, the residuals are actually more correlated than fresh data. The ML method, however, fails to adjust for this fact. As a result it uses higher degrees of freedom than what it should. The REML method is an adjustment to ML to incorporate a *correction of the degrees of freedom*. The Further Insights section at the end of this chapter provides further information on this theme. REML, like ML, belongs to the family of likelihood-based methods, and hence inherits the good qualities of the family. Thanks to the *degrees of freedom correction*, REML estimators are also unbiased. So it is the most popular estimation method for mixed effects models. This is also the default in the MIXED command.

To request this method of estimation use the following option in the ESTIMATE command line:

```
ESTIMATE /METHOD = REML
```

Some salient aspects of the REML method of estimating variance components are:

- These estimators are consistent, asymptotically efficient, and asymptotically normal under quite general assumptions.
- Asymptotic covariance matrix of the estimators is produced as a by product. It is the inverse of the information matrix.
- SYSTAT implements REML estimation using an iterative algorithm.

- REML estimation in SYSTAT depends on normality assumptions.
- REML estimators are unbiased.

Testing Hypotheses

SYSTAT can perform three different types of hypothesis tests. First, it tests each of the fixed effect coefficients for significance using the standard t-test. The square of the t-test statistic may be considered as the test statistic in an ANOVA-like F-test with a single degree of freedom in the numerator. SYSTAT reports the value of the t-statistic and also the two-sided p -value. Thus, if α_i is a fixed effect coefficient, then SYSTAT tests

$$H_0: \alpha_i = 0$$

against

$$H_1: \alpha_i \neq 0.$$

Since t-distributions are symmetric, the information can be also used for one-sided alternatives. For one-sided alternatives, the p -value is half of the reported two-sided p -value. The smaller the p -value, the more significant is the coefficient. To test at a given level of significance, say 5%, one should reject the null hypothesis if the p -value falls below 0.05.

SYSTAT also lets the user test various contrasts in three inference spaces: broad, intermediate and narrow. For a detailed real life example of hypothesis testing in different inference spaces, please see the chapter Linear Mixed Models in this manual. In this section we present only the conceptual underpinnings.

Any statistical test of hypothesis looks at the data only through the test statistic, and tries to calibrate the observed value of the statistic as *large* or *small*. This calibration is achieved by comparing the observed value with values obtained from (hypothetical) replications from a model where the null hypothesis is indeed true. As we have already mentioned in the context of standard errors, there are three different replication modes possible for a mixed model:

- We freshly randomize all the random effects (*Broad Inference Space*)
- We hold all the random effects fixed at the same levels (*Narrow Inference Space*)
- We randomize some of the random effects, and hold rest fixed (*Intermediate Inference Space*)

The hypothesis as well as the inference space are specified to SYSTAT by using three matrices F , R and D . Actually, SYSTAT tests the hypothesis $H_0: F\beta + R\gamma = D$.

Notice that this equation imposes a size restriction on the matrices F , R , and D . They must all have the same number of rows. Also, D must be a column vector. The number of columns in F must equal the number of fixed effect coefficients, while that of R must match the number of fixed effect coefficients. We introduce the three matrices one by one.

The F Matrix

Consider the model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where μ and α_i s are fixed effects and β_j 's are random. Let us assume that $i=1,2,3$ and $j=1,2,3,4$. Consider all the fixed effect parameters to be laid out in a row:

$$\mu, \alpha_1, \alpha_2, \alpha_3$$

To test the null hypothesis

$$H_0: \alpha_1 = \alpha_2$$

we rewrite it as:

$$0 \times \mu + 1 \times \alpha_1 + (-1) \times \alpha_2 + 0 \times \alpha_3 = 0.$$

Collecting the coefficients of the fixed effects parameters, we get the row vector (0, 1, -1, 0) as the F matrix. In SYSTAT we use

```
HYPOTHESIS
FMATRIX [0 1 -1 0]
TEST
```

Similarly, the F matrix (0, 1, 0, -1) tests the equality of 1 and 3. If we stack these two rows one on top of the other we get

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix}$$

Then we are testing equality of all the α_i 's. However, note that SYSTAT expects the F matrix to have independent rows. In other words, you cannot have redundant conditions on the parameters. In particular, to test equality of the α_i 's we cannot use the F matrix

$$\begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Here the three rows correspond to the three conditions: $\alpha_1 - \alpha_2 = 0$, $\alpha_1 - \alpha_3 = 0$, $\alpha_2 - \alpha_3 = 0$, of which the last condition is redundant, since it is implied by the first two.

The D Matrix

Consider testing the null hypothesis $\alpha_1 - \alpha_2 = 2$. Here we have a nonzero right hand side of the equation. SYSTAT calls the right hand side the D matrix of the hypothesis. So we shall write

```
HYPOTHESIS
  FMATRIX [0 1 -1 0]
  DMATRIX [2]
TEST
```

Obviously, the F and the D matrix must have the same number of rows.

The F matrix and the D matrix will suffice for most purposes. The resulting tests are performed in the so called broad inference space. To test hypotheses in the narrow or intermediate inference spaces we need the R-matrix, which is discussed next.

The R Matrix

In a mixed model there are two sets of random variables, the random effects and the random errors. If we pick some of the random effects and condition the model on them, then the resulting inferences are said to be done in intermediate inference space. If we condition our inference on all the random effects, then we are in the narrow inference space.

Illustrative case: Consider the model

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where $i,j=1,2,3$. We shall treat μ and α_i 's as fixed effects, and β_j 's as random.

A typical narrow inference space hypothesis here is

$$H_0: \alpha_1 - 3\alpha_2 + 2\beta_1 + 2\beta_2 + 2\beta_3 = 0.$$

Note that the hypothesis involves the random coefficients. The interpretation of this as follows: The test statistic here is to be calibrated against replications where the random effects are randomized under the constraint of H_0 . Thus, we are testing the equality of α_1 and α_2 holding the average contribution of the β_j 's fixed. This corresponds to the F matrix (0 3 -3 0), R matrix (2 2 2), and D matrix 0.

The input is:

```

HYPOTHESIS
FMATRIX [0 3 -3 0]
RMATRIX [2 2 2]
TEST

```

We have not mentioned the D matrix explicitly. So it would default to the zero matrix.

If we keep some zero entries in R, then the corresponding random coefficients are unconstrained, and we have an example of intermediate inference space. (McLean, Sanders and Stroup 1991). SYSTAT insists that all the rows of the combined matrix [F R] should be linearly independent. D should have only one column. All the three matrices default to 0.

Pairwise Comparison Tests

Consider the model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $i=1,2,3,4$, and $j=1,\dots,10$. Suppose that we have rejected the null hypothesis that α_i 's are the same. This only tells us that possibly not all the α_i 's are the same, but does not shed any light on exactly which pair(s) of α_i 's is (are) different. In this context, several (multiple) comparisons or post hoc tests (say, all pairwise comparisons) are carried out. To guard against chance conclusions of significant differences, levels of significance are adjusted when multiple comparisons are made. There are a number of methods available to make such adjustments of the individual tests in order to achieve a required overall level of significance. SYSTAT implements 6 of these:

- Bonferroni (BONF)
- Fisher's LSD (LSD)
- Tukey (TUKEY)
- Sidak (SIDAK)
- Scheffe (SCHEFFE)
- GT2 (GT2)

For more information on multiple comparisons see Chapter 1: Linear Models, "Pairwise Mean Comparisons" in *Statistics II*. The following command lines perform Bonferroni adjustment when all pairwise comparisons are made.

```
HYPOTHESIS  
  PAIRWISE X / BONF  
TEST
```

Diagnostics

Any statistical analysis makes some assumptions (in the form of a model) about the process that has generated the data to be analyzed. The model is at best an approximation to the real process. The result of the analysis is meaningful only if the

- model assumptions are correct
- model captures most of the important aspects of the data

- data does not have influential outliers (i.e., some observations that do not conform to the general pattern laid out by the model, and has the potential to distort the results.)

Any proper statistical analysis therefore must watch out for possible violations of the model assumptions. SYSTAT provides a number of ways to perform such diagnostic checks. These come in two flavors:

- Residual diagnostics
- Model selection

Residual Diagnostics

Most statistical models aim to explain the variability present in the data set by ascribing parts of it to various known causes. However, it is never possible to explain the entire variability in this way. The remaining variability is ascribed to chance. That is why we have the random error terms in (mixed) models. In these random errors we sum up all the causes of variability that we are ignorant about. This ignorance often takes the form of the assumption that the errors present in the different observations are independent and identically distributed. To perform tests of hypotheses and confidence interval estimation we also assume that these random errors follow a normal distribution with mean 0. Residual diagnostics try to check for departures from these assumptions about the random error.

The main idea behind these is essentially this: first fit the model of your choice to the data set, and obtain an approximation to the random errors by subtracting the prediction from the actual observations. These approximations are called residuals and are used as a proxy to the actual unobservable random errors. A word of caution, though: the residuals are not the actual random errors. In particular, the residuals are correlated even for a model with independent random errors. Residual diagnostic methods, therefore, should only be used as a rough check, rather than a rigorous one. However, residual checks should always be done.

In a mixed model there are two different types of residuals.

- Marginal residuals (MRESIDUALS)
- Conditional residuals (CRESIDUALS)

The names in parentheses are the terms used by SYSTAT. Marginal residuals are obtained by subtracting from the original data the prediction based on only the fixed effects. We can think of this prediction as over the population of levels of random

effects. For instance, consider the model,

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where β_j 's are random effects. Then the marginal residuals are:

$$y_{ijk} - (\hat{\mu} - \hat{\alpha}_i)$$

If we include the random effects in our prediction then we get the conditional residuals

$$y_{ijk} - (\hat{\mu} - \hat{\alpha}_i + \hat{\beta}_j)$$

You may think of the term inside parentheses as the prediction of y_{ijk} over the specific levels of random effects. Ideally both these residuals should be small and show no pattern. To check this, one should plot these against cases, and also against the covariates. The variability in the conditional residual plot will be less than that in the marginal residual plot.

Model Selection Criteria

In SYSTAT, likelihood-based model selection criteria are provided for model selection; they are:

- $AIC = -2 \text{ Log-likelihood} + 2k$
- $AIC \text{ (corrected)} = -2 \text{ Log-likelihood} + 2k + 2k(k+1)/(n-k-1)$
- $BIC = -2 \text{ Log-likelihood} + k \log(n)$

where n is the number of observations and k is the number of parameters (fixed effects, variance covariance parameters) estimated. In the REML method of estimation, AIC should be used only to compare models with the same fixed effects part. For more information on AIC and BIC see Chapter 2: Linear Models, "Variables selection" in *Statistics-II*.

Missing Observations

There are many real-life situations where some of the observations are missing from the data set. However, such incomplete data sets may be quite informative if analyzed properly. The first thing to keep in mind when dealing with missing data is the missing data mechanism: "Why are the missing observations missing?" Sometimes the observations may be missing completely at random (MCAR). For instance, some observations may get lost or corrupted during transcription. This is a random process independent of the process of interest. In some situations, on the other hand, the missing process is dependent on the data. If your data consist of the measured intensities of stars then the observations of the weaker or more distant stars are more likely to be missing. This is an example of censored data. Yet another situation is truncation. Suppose that we are observing the lifetime of electric bulbs. The bulbs are turned on at the start of the experiment, and the time when they burn out are observed. However, the experiment is conducted within a limited amount of time. We cannot report the exact lifetimes of the bulbs that continue to burn after this period is over. This is another type of missing data mechanism. The MCAR mechanism is the most popularly used assumption to cope with cases where no information is available. SYSTAT also uses this assumption.

When using likelihood-based methods like ML or REML, SYSTAT uses the Expectation-Maximization (EM) Algorithm (Dempster, Laird, and Rubin (1977)). This algorithm, as its name suggests, consists of two parts: the Expectation part and the Maximization part. In the expectation part we first pretend that we have all the observations, and compute the log-likelihood accordingly. This leads to a function of the parameters as well as the data (both observed and unobserved). Then we take the conditional expectation of this given the observed data with respect to the distribution of the missing values treated as random variables. This completes the Expectation part. In the Maximization part, we maximize this conditional expectation with respect to the parameters. This process is repeated until convergence.

Further Insights

Henderson's Mixed Model Equation

SYSTAT computes the BLUEs of the fixed effects and the BLUPs of the random effects by solving Henderson's Mixed Model Equation (MME):

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

Henderson (1953) proved that the solution to this equation maximizes the joint density of y and γ . Notice the G^{-1} in the lower right block of the coefficient matrix. Without this term the solution would be just the maximum likelihood estimator considering γ as fixed effects. Also note that in the absence of any random effects term (no Z and G) the MME reduces to the familiar system of normal equations for linear models.

The MME involves the covariance matrices G and R , which are unknown. So before we can solve the MME we need to estimate these by ML or REML.

Some Properties of BLUPs

We have mentioned earlier in this chapter that BLUPs are empirical Bayes estimators under normality assumption. Also they are a form of shrinkage estimator. We demonstrate these using the following example.

Consider the model,

$$y_i = \mu + \varepsilon_i$$

where $i=1, \dots, n$. We shall treat μ as a random effect with a $N(0, \tau^2)$ distribution, while the errors have independent $N(0, \sigma^2)$ distributions. This can be thought of as a Bayesian inference problem where the parameter μ has a $N(0, \tau^2)$ prior. Then the Bayes estimator of μ will be the posterior mean (which is the conditional expectation of μ given the data).

However, this is not an honest Bayes estimator as it involves the unknown variances. So we need to estimate them separately and plug them into the formula, producing an empirical Bayes estimator. The above formula for $\hat{\mu}$ is precisely what one would get by solving the MME, which now takes the following form:

$$(1'_n(\sigma^2 I) 1_n + \tau^{-2}) \hat{\mu} = 1'_n(\sigma^{-2} I) y$$

where 1_n denotes the n -dimensional column vector of ones. Notice that $\hat{\mu}$ is not unbiased if you consider it as an estimator of the fixed effect μ . But as a predictor of the random effect μ it is unbiased. Finally, if we consider μ as a fixed effect, then its BLUE is just \bar{y} . Thanks to the positive term $\sigma^2 \tau^{-2}$ in the denominator of the BLUP, the latter is always smaller in absolute value than \bar{y} . In other words, we can think of the BLUP as a *shrunk* version of the BLUE. This also implies that the BLUP has lower variance than the corresponding BLUE.

Why Random Effect Coefficients are Always Estimable

SYSTAT uses effects encoding for the fixed effects, but means encoding for the random effects. The need for this difference stems from the fact that, unlike the fixed effects, the random effect coefficients are always estimable (or predictable, rather!). The following example demonstrates this point.

Consider the model

$$y_i = \mu_1 + \mu_2 + \varepsilon_i$$

where $i=1, \dots, 10$. Let us assume that the ε_i 's are independent $N(0, \sigma^2)$ random variables. Now, if the μ 's are treated as fixed effects, then they are not individually estimable, since they enter the model only as their sum, which is estimable with BLUE given by \bar{y} . Any two pairs of estimators ($\hat{\mu}_1, \hat{\mu}_2$) with sum \bar{y} , would fit the data equally well, and there is nothing in the assumptions to choose one pair over another. Thus, the μ 's are not estimable here. If, on the other hand, the μ 's are random effects, and if we assume that they are independent $N(0, \tau^2)$ random variables, then certain pairs will be more likely than others. Indeed, the pair where $\hat{\mu}_1 = \hat{\mu}_2$ would be the most likely pair, and so the BLUP for each of the μ 's would be half of \bar{y} .

ML and REML

Under normality assumption, the data vector y has a multivariate normal distribution with mean vector $X\beta$ and covariance matrix $V=Z'GZ + R$. So the log-likelihood of the data is

$$(2\pi)^{(-n)/2} |V|^{(-1)/2} \exp \left[-\frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta) \right]$$

In the ML method we maximize this with respect to β , G and R . The ML estimators have asymptotic normal distribution. However, for finite sample these may be biased estimators. REML is very similar to ML except that the estimators are unbiased. Here we first fit a model containing only the fixed effects. Treat the residuals from this fit as the new data to which we shall fit the random effects. The maximum likelihood estimator obtained from this second model is the REML estimator. The interested reader may find the details in Searle, Casella, and McCulloch (1992, p.251). The following illustration may serve to clear up the distinction between ML and REML.

Illustrative case: Suppose that our data consist of y_1, \dots, y_n , which we model as

$$y_i = \mu + \varepsilon_i$$

Here μ is a fixed effect and the ε_i 's are independent $N(0, \sigma^2)$. The ML estimator of σ^2 is easily seen to be

$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

which is biased. Now let us "take out" the fixed effect estimate $\hat{\mu} = \bar{y}$ to get the residuals $y_i - \bar{y}$. If we treat these as our new data, and compute the maximum likelihood estimator of σ^2 , then we shall obtain the REML estimator

$$\hat{\sigma}_{REML}^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

which is unbiased.

References

- *Beckman, R.J., Nachtsheim, C.J., and Cook, D.J. (1987). Diagnostics for mixed model analysis of variance. *Technometrics*, 29, 413-426.
- *Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). *Regression diagnostics; Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.
- *Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.
- *Brown, H. and Prescott, R. (1999). *Applied Mixed Models in Medicine*. New York: John Wiley & Sons.
- *Brownlee, K.A. (1960). *Statistical Theory and Methodology in Science and Engineering*. New York: John Wiley & Sons.
- *Burdick, R.K. and Graybill, F.A. (1992). *Confidence intervals on variance components*. New York: Marcel Dekker.
- *Christensen, R., Pearson, L.M., and Johnson, W. (1992). Case-deletion diagnostics for mixed models. *Technometrics*, 34, 38-45.
- *Crowder, M.J. and Hand, D.J. (1990). *Analysis of repeated measures*. New York: Chapman and Hall.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B*, 39, 1-38.
- *Diggle, P.J. (1990). *Time series: A biostatistical introduction*. Oxford: Oxford University Press.
- *Diggle, P.J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-93.
- *Everitt, B.S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, 44, 113-135.
- *Fellner, W.H. (1986). Robust estimation of variance components. *Technometrics*, 28, 51-60.
- *Hand, D.J., Daly, F., McConway, K., and Lunn, D. (1994). *A handbook of small data sets*. London: Chapman Hall.
- *Hartley, H.O. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54, 93-108.
- *Harville, D.A. (1990). BLUP (Best Linear Unbiased Prediction) and beyond. *Advances in Statistical Methods for Genetic Improvement of Livestock*: Gianola, D., Hammond, K. (eds.). pp.239-276. Berlin: Springer-Verlag.
- *Harville, D.A. and Jeske, D.R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.

- Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.
- Henderson, C.R., Kempthorne, O., Searle, S.R., and von Krosig, C.N.. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics*, 15, 192-218.
- *Hocking, R.R. (2003). *Methods and applications of linear models*. New York: John Wiley & Sons.
- *Kuehl, R.O. (2000). *Design of experiments: Statistical principles of research design and analysis*. New York: Duxbury Thomson Learning.
- McLean, R.A., Sanders, W.L., and Stroup, W.W. (1991). A unified approach to mixed linear models. *The American Statistician*, 45, 54-64.
- Mickey, R. M., Dunn, O. J., and Clark, V. A. (2004). *Applied statistics: Analysis of variance and regression*. New York: John Wiley & Sons.
- Milliken, G.A. and Johnson, D.E. (1992). *Analysis of messy data, Volume I: Designed experiments*. London: Chapman and Hall.
- Netmaster Statistics Courses. Available at:
<http://www.dina.kvl.dk/~per/Netmaster/courses/st113/Data/datafiles/planks.txt>.
- *Ostle, B. and Malone, L.C. (1988) *Statistics in research*, 4th ed . Ames, Iowa: State University Press.
- Rao, C.R. (1971). Estimation of variance and covariance components. *Journal of Multivariate Analysis*, 1, 257-275.
- Searle, S.R., Casella, G., and McCulloch, C.E. (1992). *Variance components*. New York: John Wiley & Sons.
- *Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.
- *Wolfinger, R.D., Tobias, R.D., and Sall, J. (1994). Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM Journal on Scientific Computing*, 15(6), 1294-1310.

(* indicates additional reference.)

Variance Components Models

Arnab Chakraborty, Ravindra Jore, Sourov Ghosh, and K. Raghavendra Rao

Variance Components (VC) can carry out estimation and hypothesis tests in a variance components model for both balanced and unbalanced data. A variance components model can have any number of fixed and/or random effects, including interactions (crossed effects) and nestings (nested effects). Both categorical and continuous variables are allowed as predictor variables. Thus VC can be used to fit mixed regression as well as mixed ANOVA models. The models handled by VC constitute a subclass of those handled by MIXED, which allows more general covariance structures for the random effects and the random error. The subclass of models dealt with by VC is arguably the most frequently used type of linear mixed models.

Statistical Background

A variance components model is a mixed linear model of the form

$$y = X\beta + Z_1\gamma_1 + \dots + Z_p\gamma_p + \varepsilon,$$

where y is the data vector, X and Z_i 's are known matrices (either design matrices or covariate matrices), β is the vector of fixed effects, each γ_i is a vector of random effects, and ε is the random error vector. Here y is a random vector, whose randomness comes partly from the random vector γ_i and partly from ε . We assume that the random vectors γ_i and ε have independent Gaussian distributions with zero mean and covariance matrices of the form

$$\text{Var}(\varepsilon) = \sigma_0^2 I, \text{Var}(\gamma_i) = \sigma_i^2 I$$

where I is an identity matrix of appropriate order. Here each γ_i consists of the random coefficients for one random effect. The variance of the distribution may be different for the different effects. (SYSTAT provides the option to specify a common variance parameter for multiple effects.) Over and above the usual estimation techniques for general linear mixed models, VC offers some extra estimation techniques specially applicable for this subclass. Unlike the general methods (ML and REML) the special methods (MIVQUE0 and ANOVA: TYPE1, TYPE2, and TYPE3) are non-iterative and require less computation. For details of these methods, please refer to chapter on "Introduction to Linear Mixed Models" on page 251, Statistics II.

SYSTAT reports the BLUE's of the fixed effects and BLUP's of the random effects, as well as estimates of the variance parameters. Each estimation or prediction is accompanied with its standard error, two-sided 95% confidence interval, and a significance test.

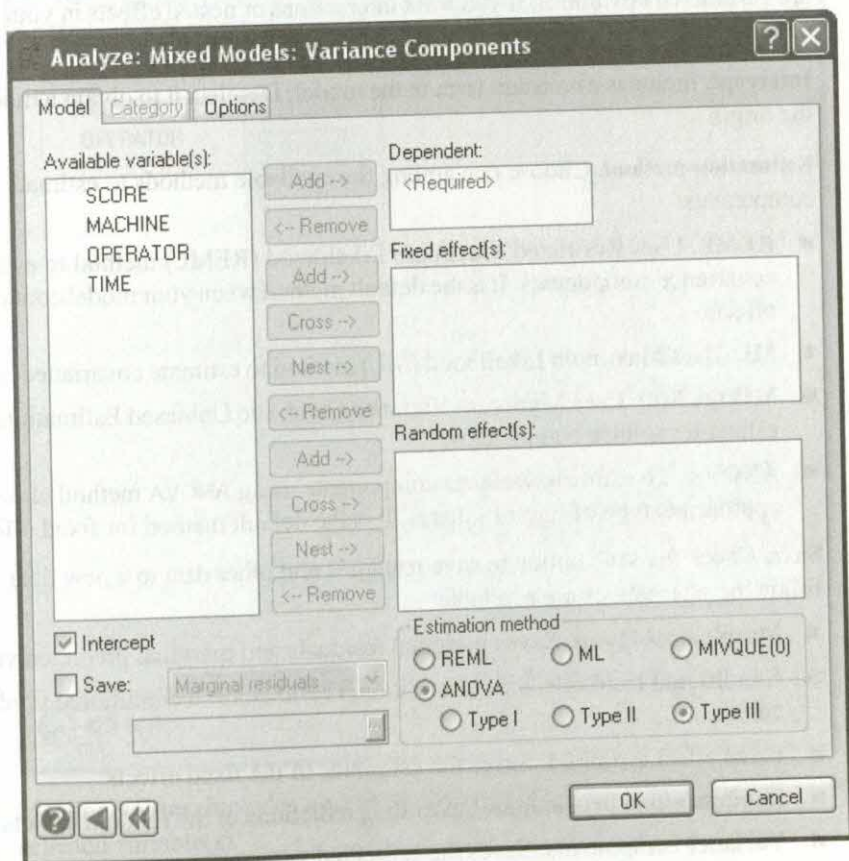
For each model you fit in VC, SYSTAT reports log-likelihood (even if you are not using a likelihood-based estimation method like ML or REML), Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), and Akaike Information Criterion Corrected (AICC) for assessing the fit of the model.

Variance Components in SYSTAT

Model Estimation (in VC)

To fit a Variance Components model, from the menus choose:

Analyze
Mixed Models
Variance Components...



To specify a variance components model and the method to estimate/predict the effects and variance components, use the available options.

Dependent. Dependent is the variable you want to model. Dependent variable should be a continuous numeric variable.

Fixed effect(s). Select one or more continuous or categorical (grouping) variables which you treat as fixed effects. Fixed effects that are not denoted as categorical are considered covariates. If you want crossed or nested effects in your model, you need to build these components using Cross and Nest buttons.

Random effect(s). Select one or more continuous or categorical (grouping) variables which you treat as random effects. Random effects that are not denoted as categorical are considered covariates. If you want interactions or nested effects in your model, you need to build these components using Cross and Nest buttons.

Intercept. Includes a constant term in the model. Deselect it to obtain a model through the origin.

Estimation method. Choose one among the available methods to estimate variance components:

- **REML.** Uses Restricted Maximum Likelihood (REML) method to estimate covariance components. It is the default method when your model contains random effects.
- **ML.** Uses Maximum Likelihood (ML) method to estimate covariance components.
- **MIVQUE(0).** Uses Minimum Variance Quadratic Unbiased Estimation method to estimate variance components.
- **ANOVA.** To estimate variance components using ANOVA method choose appropriate type of sum of squares. It is the default method for fixed effects model.

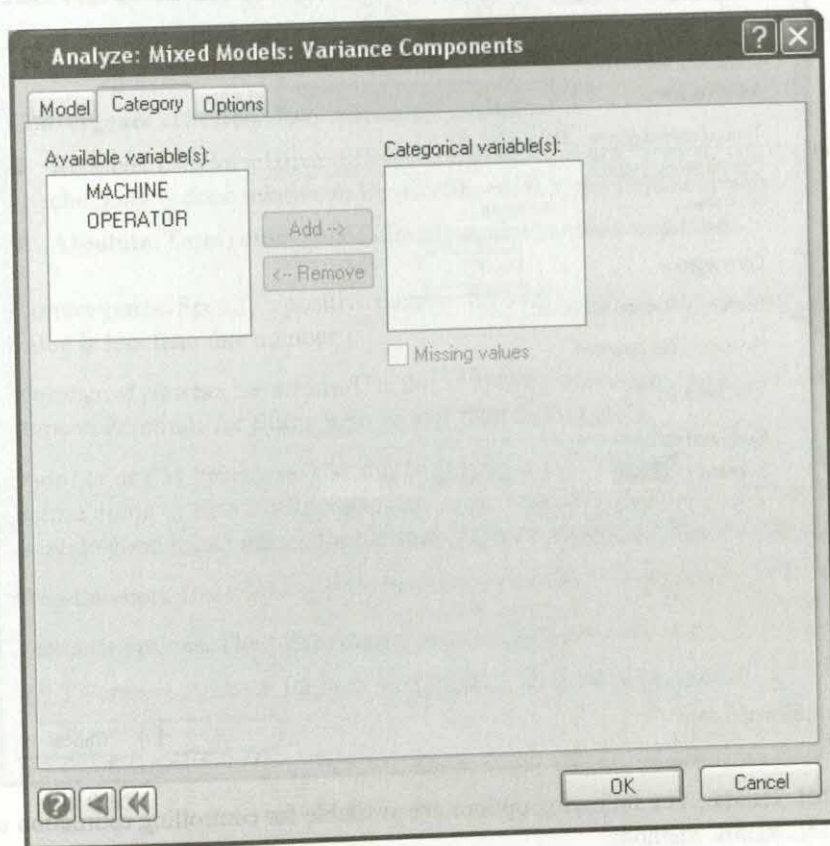
Save. Check the save option to save residuals and other data to a new data file. The following alternatives are available:

- **Marginal residuals.** Saves marginal residuals and marginal predicted values.
- **Conditional residuals.** Saves conditional residuals and conditional predicted values.
- **Fixed effect estimates.** Saves the estimates of the fixed effects.
- **Random effect predictions.** Saves the predictions of the random effects.
- **Variance components.** Saves the estimated variance components.
- **Standard errors of fixed effects.** Saves standard errors of the fixed effect estimates.
- **Residuals/data.** Saves marginal and conditional residuals along with all the variables in the working data file.

- **Model.** Saves marginal residuals, response variable, and the design matrices.

Category

To specify categorical variables, click the Category tab. Select at least one fixed or random effect in Model tab other than intercept to activate this tab.



Missing values. Includes a separate category for cases with a missing value for the selected variable(s).

Options

Use Options tab to specify computational controls for ML or REML method of estimation.

The screenshot shows a dialog box titled "Analyze: Mixed Models: Variance Components" with three tabs: "Model", "Category", and "Options". The "Options" tab is selected. The dialog contains the following fields and controls:

- ML/REML** section:
 - Initial values:** An empty text box.
 - Type of convergence:** A dropdown menu with "Hessian" selected.
 - Convergence criterion:** Two radio buttons: "Relative" (selected) and "Absolute".
 - Convergence:** A text box containing "1e-008".
 - Number of Newton iterations:** A text box containing "20".
 - Number of EM iterations:** A text box containing "5".
 - Step-halvings:** A text box containing "50".
- Estimation options** section:
 - Tolerance:** A text box containing "1e-012".
 - Confidence:** A text box containing "0.95".

At the bottom of the dialog, there are navigation buttons (back, forward, and a question mark) and "OK" and "Cancel" buttons.

ML/REML. The following options are available for controlling estimation using ML/REML methods:

- **Initial values.** Use this option to provide initial values for variance components. Specify values for each component in the order the effects appear in your model. Separate the values with commas or blanks. Do not specify initial values for some of the parameters and leave blanks for others. If you do, SYSTAT computes initial values for all variance components. Make sure that the initial values are positive and are in a reasonable range.

Type of convergence. Check one of the following options to check convergence.

Three types of convergence checks are available:

- **Hessian.** Uses a quadratic form $g' H^{-1} g$ where g is the gradient vector and H is the hessian matrix.
- **Likelihood.** Uses the difference between log-likelihood at current iteration and the log-likelihood at last iteration.
- **Parameter.** Uses maximum of absolute differences between parameter estimates at current iteration and parameter estimates at last iteration.

Convergence criterion. Two criteria are available:

- **Relative.** Checks relative difference for convergence. That is, convergence checking is done relative to log-likelihood. It is the default option.
- **Absolute.** Tests convergence directly against a value specified.

Convergence. Specify a positive number. SYSTAT stops iterations when convergence value is less than this number.

Number of Newton iterations. Use this to specify maximum number of Newton-Raphson iterations for fitting your model. The default is 20.

Number of EM iterations. Use this to specify maximum number of EM iterations before going to Newton-Raphson iterations. Sufficient number of EM iterations provide good initial values for Newton-Raphson iterations. The default is 5.

Step-halvings. Use this to specify maximum number of step halvings. The default is 50.

Estimate options. The following estimate options are available:

- **Tolerance.** A check for near singularity. Use Tolerance to guard against this singularity problem.
- **Confidence.** Specifies the confidence coefficient for testing purposes. The default is 0.95.

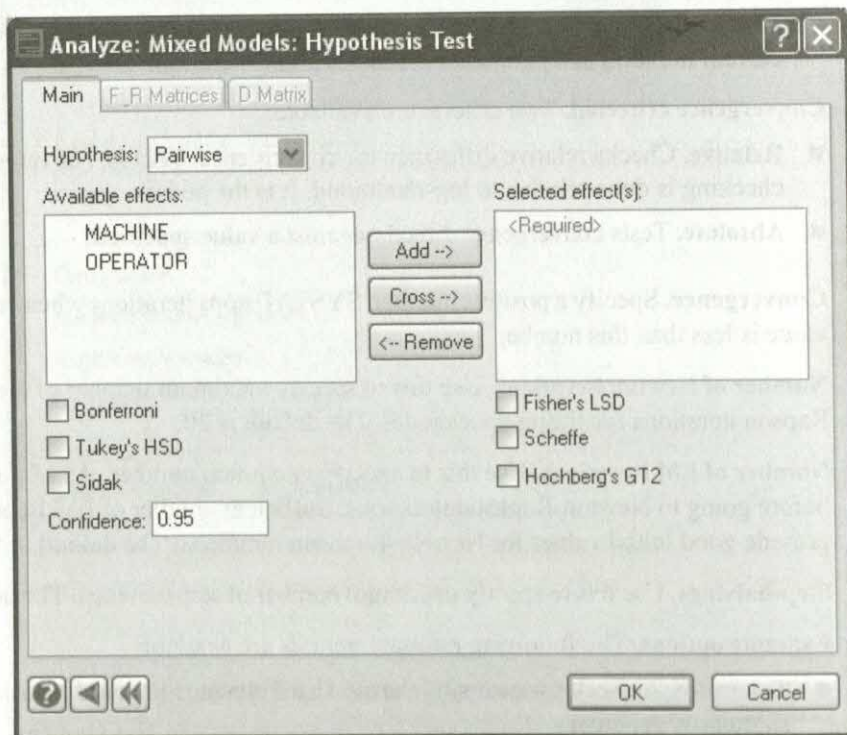
Hypothesis Test

To test hypotheses, from the menus choose:

Analyze

Mixed Models

Hypothesis Test...



You can customize the hypothesis to be tested. Contrasts can be defined across the categories of a grouping factor:

Hypothesis. Select the type of hypothesis. The following choices are available:

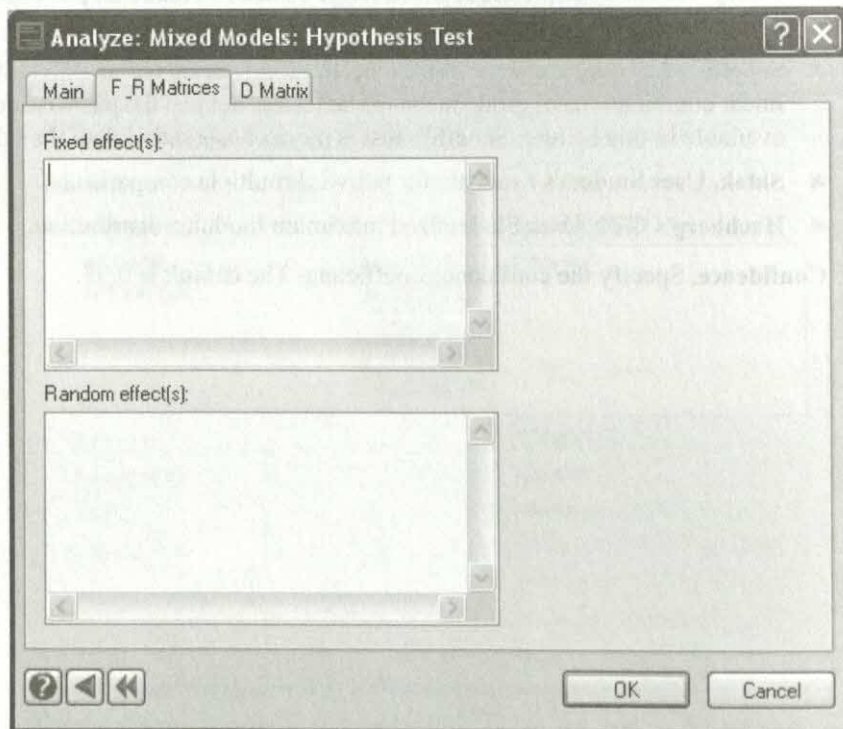
- **Pairwise.** Compare pairs of groups to determine which pairs differ.
- **F and R Matrices.** Tests the hypotheses corresponding to the F and R Matrices tab.

Adjustment method. The following options are available to compute *p-value* adjusted for multiple comparisons:

- **Bonferroni.** Uses Student's t statistics. It sets the family-wise error rate as $(1 - \text{Confidence}) / (\text{Total number of comparisons})$.
 - **Fisher's LSD.** Equivalent to multiple t tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.
 - **Tukey's HSD.** Uses the Studentized range statistic to make all pairwise comparisons. This is the default.
 - **Scheffé.** The significance level of Scheffé's test is designed to allow all possible linear combinations of group means to be tested, not just the pairwise comparisons available in this feature. Scheffé's test is more conservative than the other tests.
 - **Sidak.** Uses Student's t statistic for pairwise multiple comparisons.
 - **Hochberg's GT2.** Uses Studentized maximum modulus distribution.
- Confidence.** Specify the confidence coefficient. The default is 0.95.

F and R Matrices

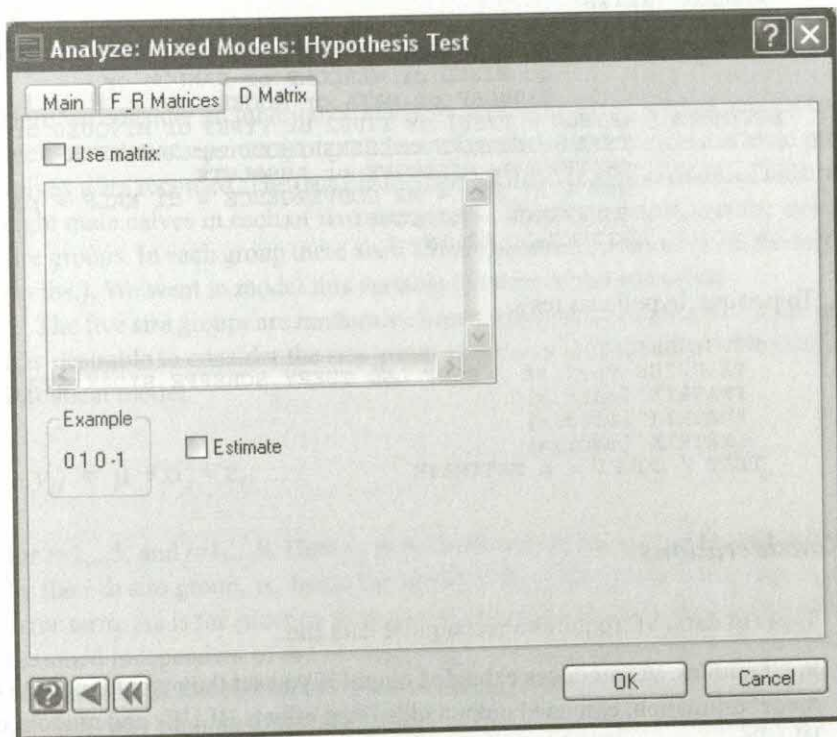
F and **R** are the matrices of linear weights contrasting the coefficient estimates for fixed and random effects respectively. You can write your hypothesis in terms of the **F** and **R** matrices.



- **Fixed effect(s).** Specify as many numbers as the dimension of your beta vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.
- **Random effect(s).** Specify as many numbers as dimension of your gamma vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.

D Matrix

D is a null hypothesis vector (by default null vector). The **D** vector, if you use it, must have the same number of rows as the **F** or **R** matrices. To specify a different **D** Matrix, click the D Matrix tab in the Analyze: Mixed Model: Hypothesis Test dialog box.



Specify a vector of dimension same as the number of rows in F and R matrices.

Estimate. Check this option for testing significance of each contrast (row) of F and R matrices individually. This test reports estimate of the estimable linear parametric function, its standard error and the corresponding t-test.

Using Commands

First, specify your data with USE filename. Continue with:

VC

```

RESET
MODEL depvar = INTERCEPT + varlist
RANDOM varlist
CATEGORY varlist
SAVE filename / MRESIDUALS or CRESIDUALS
                  or FIXED or VARCOMP or RANDOM or
                  SERRORF or DATA or MODEL
ESTIMATE / METHOD = TYPE1 or TYPE2 or TYPE3 or MIVQUE0 or ML or REML
            TYPE = HESSIAN or LIKELIHOOD or PARAMETERS
            CRITERION= RELATIVE or ABSOLUTE
            NEM = n1 NNR = n2 CONVERGENCE = d1 HALF = n3
            TOLERANCE = d2 CONFI = n4
            GSTART=11,12...1Nv.]

```

To perform hypothesis tests:

```

HYPOTHESIS
PAIRWISE varlist / BONF LSD TUKEY SCHEFFE SIDAK GT2
FMATRIX [matrix]
RMATRIX [matrix]
DMATRIX [matrix]
TEST / CONFI = n ESTIMATE

```

Usage Considerations

Types of data. VC requires a rectangular data file.

Print options. VC produces extended output if you set the output length to LONG. For model estimation, extended output adds fixed effects BLUEs and random effects BLUPs.

Quick Graph. VC produces a Quick Graph of marginal residuals versus marginal predicted values.

Saving files. Several sets of output can be saved to a file. The actual contents of the saved file depend on the analysis. Files may include estimated regression coefficients, model variables, residuals, and predicted values.

BY groups. VC analyzes data by groups.

Case frequencies. VC uses the FREQUENCY variable, if present, to duplicate cases.

Case weights. VC uses the values of any WEIGHT variables to weight each case.

Examples

Example 1

Getting Acquainted with the Output Layout

Here we consider an inheritance study on beef animals of several sire groups (males) each mated to a separate group of dams (females). Birth weights of male progeny calves were recorded. The datafile *KUEHL* (Kuehl, 2000), consists of birth weights of eight male calves in each of five sire groups. The first column lists the indices of the sire groups. In each group there are 8 sires. The second column gives the birth weights (in lbs.). We want to model this variable in terms of the sire effect.

The five sire groups are randomly chosen from a large population of sire groups. So, it is desirable to consider the sire group effect as a random effect. We shall use the statistical model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

for $i=1, \dots, 5$, and $j=1, \dots, 8$. Here y_{ij} is the birth weight (in lbs.) of the j -th calf produced by the i -th sire group, α_i being the random effect due to i -th sire group, ε_{ij} being the error term. As is the practice with any mixed effects model, the random effect is assumed independent of the random error. It is to be noted that there are two variance components in this model, one for sire and one for error term. The input for analyzing this one-way random effect model is:

```
USE KUEHL
VC
CATEGORY SIRE
MODEL BIRTHW = INTERCEPT
RANDOM SIRE
ESTIMATE
```

The output is:

Categorical values encountered during processing are

Variables	Levels
SIRE (5 levels)	177.000 200.000 201.000 202.000 203.000

Dependent Variable : BIRTHW
 Fixed Covariate(s) : Intercept
 Random Factor(s) : SIRE
 Estimation Method : Residual or Restricted Maximum Likelihood (REML)

Dimensions

Covariance Parameters : 2
 Columns in X : 1
 Columns in Z : 5
 No. of Observations : 40

There are two variance components. So the number of variance parameters is 2. X is the design matrix for the fixed part. Here it consists of only the intercept term. So it consists of just a single column. The five sire groups account for the five columns of the random effects design matrix Z.

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		359.579	
1	ECME	358.651	0.003
2	ECME	358.372	0.001
3	ECME	358.277	0.000
4	ECME	358.242	0.000
5	ECME	358.228	0.000
6	NR	358.217	0.000
7	NR	358.217	0.000
8	NR	358.217	0.000

This table contains information of a somewhat sophisticated nature. The estimation method used is REML, which is an iterative method. This table reports the convergence status of the algorithm. You would rarely need to look at it except to check if the estimation has converged at all. For a convergent process the final entry in the last column would be zero, as here. Here the convergence has been attained after 12 steps. The ECME steps can be thought of as part of initialization, which is followed by the main Newton-Raphson (NR) steps. The convergence criterion is minus twice log-

likelihood, the values of which are reported in the third column. It is possible to change this default criterion, by using options for the ESTIMATE command in SYSTAT.

Fit Statistics

```
Final L-L      : -179.108
-2L-L         : 358.217
AIC           : 362.217
AIC(Corrected) : 362.550
BIC           : 365.544
```

As discussed earlier smaller values indicate a more parsimonious fit.

Estimates of Covariance Components

Random Effect	Description	Estimate
SIRE	Variance Parameter	116.749
Error variance	Variance Parameter	463.793

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	82.550	5.911	4	13.965	0.000

Confidence Intervals of Fixed Effects Estimates

Effect	Estimate	95.00% Confidence Interval	
		Lower	Upper
Intercept	82.550	66.137	98.963

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
SIRE	177	0.718	7.372	35	0.097	0.923
	200	9.990	7.372	35	1.355	0.184
	201	-12.980	7.372	35	-1.761	0.087
	202	-3.374	7.372	35	-0.458	0.650
	203	5.646	7.372	35	0.766	0.449

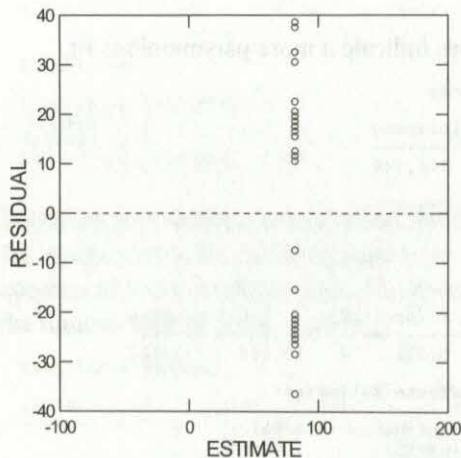
Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
SIRE	177	0.718	-14.247	15.683
	200	9.990	-4.976	24.955
	201	-12.980	-27.945	1.985
	202	-3.374	-18.339	11.591
	203	5.646	-9.319	20.611

These are the estimates and predictions of the fixed and random effects. Note that SYSTAT provides point estimates (or predictions) as well as confidence intervals. A two-tailed t-test is also reported for each effect. The degrees of freedom (df) reported for each effect is the error degrees of freedom from the ANOVA table. The confidence interval gives an idea about the precision of the estimate. The 95% confidence interval

for the intercept term, for example, is (66.137, 98.963). This roughly means that the actual intercept value (which is unknown) lies within this range with 95% chance. Similar explanation applies to the random effect confidence intervals.

Plot of Residuals vs Predicted Values



SYSTAT always produces this Quick Graph by default. The vertical axis shows the residuals and the horizontal axis shows the estimate. The estimate is obtained by considering only the fixed effect(s). In this example the only fixed effect is the intercept term. So, all the estimates have the same value. This is seen easily from the plot, all the points are along the same vertical line.

The *p*-value of the intercept is 0.000. So the intercept term is significant at, say, 0.05 level. This is of course no great news, since all that it says is that the calves weigh significantly different from 0 at birth. A more important piece of news lies in the *p*-values for the random effects. All of them are pretty large (more than 0.05, say) and so none of the sire effects appear significant. Thus we conclude that the birth weight of a calf does not really depend significantly on the group of its dad.

Example 2

A Model with Interaction

This example illustrates how the VC command handles interaction effects. An experiment, described in Milliken and Johnson (1992) was conducted by a company to compare performances of three different brands of machines when operated by the company's own personnel. Six employees were selected at random and each of them had to operate each machine three different times. The file MACHINE1 contains these data. The data set consists of overall scores that take into account both the quality and quantity of output

We have a two-way treatment structure here. However, since certain operators may find certain brands of machines more (or less) difficult to use, we cannot a priori ignore the possibility of an interaction between the operator and machine effects. So our model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the score of the j -th operator operating the i -th machine at the k -th time point. The operator effect β_j is assumed random, since the operators were selected at random from among the employees. The machine effects α_i are fixed.

The input is:

```
USE MACHINE1
VC
CATEGORY MACHINE OPERATOR
MODEL SCORE = INTERCEPT + MACHINE
RANDOM OPERATOR + MACHINE * OPERATOR
ESTIMATE
```

The following are selections from the output:

Categorical values encountered during processing are

Variables	Levels					
MACHINE (3 levels)	1.000	2.000	3.000			
OPERATOR (6 levels)	1.000	2.000	3.000	4.000	5.000	6.000
Dependent Variable :	SCORE					
Fixed Factor(s) :	MACHINE					

Fixed Covariate(s) : Intercept
 Random Factor(s) : OPERATOR, MACHINE*OPERATOR
 Estimation Method : Residual or Restricted Maximum Likelihood (REML)

Dimensions

Covariance Parameters : 3
 Columns in X : 4
 Columns in Z : 24
 No. of Observations : 54

The 4 columns of X are from the intercept and the 3 machines. The 6 operators and 6 times 3 interactions account for the 24 columns of Z.

Estimates of Covariance Components

Random Effect	Description	Estimate
OPERATOR	Variance Parameter	22.858
MACHINE*OPERATOR	Variance Parameter	13.909
Error variance	Variance Parameter	0.925

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		66.272	2.486	5	26.660	0.000
MACHINE	1	-13.917	2.177	10	-6.393	0.000
	2	-5.950	2.177	10	-2.733	0.021
	3	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		66.272	60.733	71.811
MACHINE	1	-13.917	-18.767	-9.066
	2	-5.950	-10.801	-1.099
	3	0.000	.	.

Notice that the estimate for the last machine is 0. Actually this is a statistical artifact to avoid non-estimability problems. It is not possible to estimate all the three α_i 's together. We need to put some extra condition. SYSTAT imposes the condition that the

last α_i equals 0. The same assumption is reflected in the confidence intervals also that are reported next.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
OPERATOR	1	1.045	2.661	36	0.393	0.697
	2	-1.376	2.661	36	-0.517	0.608
	3	5.361	2.661	36	2.015	0.051
	4	-0.060	2.661	36	-0.022	0.982
	5	2.545	2.661	36	0.956	0.345
	6	-7.514	2.661	36	-2.824	0.008
MACHINE*OPERATOR	1*1	-0.750	2.388	36	-0.314	0.755
	1*2	1.553	2.388	36	0.650	0.520
	1*3	1.778	2.388	36	0.745	0.461
	1*4	-1.039	2.388	36	-0.435	0.666
	1*5	-3.457	2.388	36	-1.448	0.156
	1*6	1.916	2.388	36	0.803	0.427
	2*1	1.500	2.388	36	0.628	0.534
	2*2	0.607	2.388	36	0.254	0.801
	2*3	2.299	2.388	36	0.963	0.342
	2*4	2.417	2.388	36	1.012	0.318
	2*5	2.152	2.388	36	0.901	0.373
	2*6	-8.976	2.388	36	-3.759	0.001
	3*1	-0.114	2.388	36	-0.048	0.962
	3*2	-2.997	2.388	36	-1.255	0.218
	3*3	-0.815	2.388	36	-0.341	0.735
	3*4	-1.414	2.388	36	-0.592	0.557
	3*5	2.853	2.388	36	1.195	0.240
	3*6	2.487	2.388	36	1.042	0.305

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
OPERATOR	1	1.045	-4.352	6.441
	2	-1.376	-6.773	4.021
	3	5.361	-0.036	10.758
	4	-0.060	-5.457	5.337
	5	2.545	-2.852	7.941
	6	-7.514	-12.911	-2.118
MACHINE*OPERATOR	1*1	-0.750	-5.592	4.092
	1*2	1.553	-3.290	6.395
	1*3	1.778	-3.065	6.620
	1*4	-1.039	-5.882	3.803
	1*5	-3.457	-8.299	1.385
	1*6	1.916	-2.926	6.758
	2*1	1.500	-3.342	6.342
	2*2	0.607	-4.235	5.449
	2*3	2.299	-2.543	7.142
	2*4	2.417	-2.425	7.260
	2*5	2.152	-2.690	6.994
	2*6	-8.976	-13.818	-4.134
	3*1	-0.114	-4.956	4.728
	3*2	-2.997	-7.839	1.846
	3*3	-0.815	-5.657	4.027
	3*4	-1.414	-6.257	3.428
	3*5	2.853	-1.989	7.695
	3*6	2.487	-2.355	7.329

A quick glance through the *p-value* column tells us which of the random effects are significant. Smaller *p-values* are more significant. If we are using level 0.05, then we

should look out for *p-values* smaller than 0.05. However, there is a caveat. When looking for significant effects we must always start from the higher order effects first. In this example, these are the interactions. Looking for significant lower order effects make sense only when higher order effects are all insignificant. In our example, however, there is significant interaction between machine2 and operator6. Judging from the large negative value of the t-statistic, operator6 was having some real difficulty with machine2.

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
MACHINE	2	10	20.576	0.000

This ANOVA table is for the fixed effects (i.e., machines in our example). It tries to tell us if the machines are significantly different. But wait! We have already found the presence of significant interaction. So we must not jump to the conclusion that the machines differ significantly just by looking at the low *p-value*. The apparent difference between the machines might very well be caused by operator6 messing up with machine2. (Remember the significant interaction term for this pair?) A good analyst should first investigate more carefully the significant interaction terms before blindly testing main effects.

Notice that this model does not take time into account. However, it may happen that a machine behaves differently when run for the first time than when it is run next. If we have reason to suspect this, then we should introduce an interaction term between machine and time.

The input is:

```
USE MACHINE1
VC
CATEGORY MACHINE OPERATOR
MODEL SCORE = INTERCEPT + MACHINE
RANDOM OPERATOR + MACHINE * OPERATOR + MACHINE * TIME
ESTIMATE
```

We have made the new interaction term a random effect assuming that the times were chosen at random. This time we shall not present the entire output (which is rather long). Instead, we shall show only the relevant portion, viz., the interaction effects with time.

The following are selections from the output:

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
OPERATOR	1	1.045	2.661	30	0.393	0.697
	2	-1.376	2.661	30	-0.517	0.609
	3	5.361	2.661	30	2.015	0.053
	4	-0.060	2.661	30	-0.022	0.982
	5	2.545	2.661	30	0.956	0.347
	6	-7.514	2.661	30	-2.824	0.008
MACHINE*OPERATOR	1*1	-0.750	2.388	30	-0.314	0.756
	1*2	1.553	2.388	30	0.650	0.520
	1*3	1.778	2.388	30	0.745	0.462
	1*4	-1.039	2.388	30	-0.435	0.666
	1*5	-3.457	2.388	30	-1.448	0.158
	1*6	1.916	2.388	30	0.803	0.429
	2*1	1.500	2.388	30	0.628	0.535
	2*2	0.607	2.388	30	0.254	0.801
	2*3	2.299	2.388	30	0.963	0.343
	2*4	2.417	2.388	30	1.012	0.319
	2*5	2.152	2.388	30	0.901	0.375
	2*6	-8.976	2.388	30	-3.759	0.001
	3*1	-0.114	2.388	30	-0.048	0.962
	3*2	-2.997	2.388	30	-1.255	0.219
	3*3	-0.815	2.388	30	-0.341	0.735
	3*4	-1.414	2.388	30	-0.592	0.558
	3*5	2.853	2.388	30	1.195	0.241
	3*6	2.487	2.388	30	1.042	0.306
MACHINE*TIME	1*1	0.000	0.008	30	-0.016	0.987
	1*2	0.000	0.008	30	0.005	0.996
	1*3	0.000	0.008	30	0.011	0.991
	2*1	0.000	0.008	30	0.001	0.999
	2*2	0.000	0.008	30	-0.013	0.990
	2*3	0.000	0.008	30	0.011	0.991
	3*1	0.000	0.008	30	0.003	0.998
	3*2	0.000	0.008	30	0.012	0.990
	3*3	0.000	0.008	30	-0.015	0.988

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
OPERATOR	1	1.045	-4.390	6.479
	2	-1.376	-6.810	4.059
	3	5.361	-0.074	10.795
	4	-0.060	-5.494	5.375
	5	2.545	-2.890	7.979
	6	-7.514	-12.949	-2.080
MACHINE*OPERATOR	1*1	-0.750	-5.626	4.126
	1*2	1.553	-3.323	6.429
	1*3	1.778	-3.098	6.654
	1*4	-1.039	-5.915	3.837
	1*5	-3.457	-8.333	1.419
	1*6	1.916	-2.960	6.792
	2*1	1.500	-3.376	6.376
	2*2	0.607	-4.269	5.483
	2*3	2.299	-2.577	7.175
	2*4	2.417	-2.459	7.293
	2*5	2.152	-2.724	7.028
	2*6	-8.976	-13.852	-4.100
	3*1	-0.114	-4.990	4.762
	3*2	-2.997	-7.873	1.879
	3*3	-0.815	-5.691	4.061
	3*4	-1.414	-6.290	3.462

	3*5	2.853	-2.023	7.729
	3*6	2.487	-2.389	7.363
<hr/>				
MACHINE*TIME	1*1	0.000	-0.016	0.016
	1*2	0.000	-0.016	0.016
	1*3	0.000	-0.016	0.016
	2*1	0.000	-0.016	0.016
	2*2	0.000	-0.016	0.016
	2*3	0.000	-0.016	0.016
	3*1	0.000	-0.016	0.016
	3*2	0.000	-0.016	0.016
	3*3	0.000	-0.016	0.016

Our interest lies in whether the interaction of machine with time is significant or not. Well, judging by the high *p-values* they are not. So we can indeed remain happy with the first model. For a more sophisticated analysis of the same data set please see the next chapter.

You might be tempted to put all possible interaction terms in your model to safeguard against overlooking any potential interactions. However, remember that interaction terms introduce more parameters, which eat up degrees of freedom. So the more interaction terms you introduce, the less degrees of freedom are left for estimating the variance components, resulting in less precise estimates.

Example 3

Nested Effects

This is an example where one effect is nested inside another, i.e., the levels of one effect have different meanings within the levels of another effect.

In this data set, from Kuehl (2000), our interest lies in comparing two standard pesticide methods. In particular, we want to find out if the amount of residue left on cotton plant leaves is the same for the two methods, which we shall call methods 1 and 2. To test this 6 batches of plants were sampled from the field. Two plants were used in the experiment from each batch. Thus, there were 12 plants in the experiment. The plants inside each batch were from the same field plot. Method 1 was applied to 3 randomly selected batches, and the remaining 3 batches were given Method 2. The amount of residue on leaves was measured after a specified amount of time for each of the 12 plants. Data are in *PESTRESIDUE* file.

We shall fit the model

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

where $i=1, 2, j=1, 2, 3$ and $k=1, 2$. Here y_{ijk} is the measurement for the k -th plant in the j -th batch under i -th method.

The input is

```

USE PESTRESIDUE
VC
CATEGORY METHOD BATCH
MODEL Y = INTERCEPT + METHOD
RANDOM BATCH (METHOD)
ESTIMATE

```

The following are selections from the output:

Categorical values encountered during processing are

Variables	Levels					
METHOD (2 levels)	1.000	2.000				
BATCH (6 levels)	1.000	2.000	3.000	4.000	5.000	6.000

Dependent Variable : Y
 Fixed Factor(s) : METHOD
 Fixed Covariate(s) : Intercept
 Random Factor(s) : BATCH(METHOD)
 Estimation Method : Residual or Restricted Maximum Likelihood (REML)

Dimensions

```

Covariance Parameters : 4
Columns in X          : 3
Columns in Z          : 6
No. of Observations  : 12

```

The three columns in X are due to the intercept and the two methods. There are 3 times 2 batch (method) nested terms giving rise to the 6 columns of Z.

Fit Statistics

```

Final L-L      : -38.503
-2L-L         : 77.005
AIC            : 81.005
AIC(Corrected) : 82.720
BIC            : 81.610

```

Estimates of Covariance Components

Random Effect	Description	Estimate
BATCH(METHOD)	Variance Parameter	67.500
Error variance	Variance Parameter	55.083

Notice that random variation present in the data originates more from the difference of the batches within the methods than random errors (e.g., measurement errors, differences among the plants within the batches etc).

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		69.833	5.629	4	12.407	0.000
METHOD	1	50.167	7.960	4	6.302	0.003
	2	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		69.833	54.206	85.461
METHOD	1	50.167	28.066	72.267
	2	0.000	.	.

The small *p-values* (less than 0.05, say) indicate significant terms. The dots in the last row are because of the estimability condition imposed on the fixed effect coefficients by SYSTAT: the last coefficient of each fixed main effect is assumed to be 0. The small *p-value* for the first method tells us that the two methods indeed differ significantly. However, before believing the *p-values* we must check the higher order terms (the nested terms, in this example). This is done next.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
BATCH (METHOD)	1 (1)	-3.551	5.962	6	-0.596	0.573
	2 (1)	-7.102	5.962	6	-1.191	0.279
	3 (1)	10.653	5.962	6	1.787	0.124
	4 (2)	0.829	5.962	6	0.139	0.894
	5 (2)	2.249	5.962	6	0.377	0.719
	6 (2)	-3.078	5.962	6	-0.516	0.624

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
BATCH (METHOD)	1 (1)	-3.551	-18.139	11.036
	2 (1)	-7.102	-21.690	7.485
	3 (1)	10.653	-3.934	25.241
	4 (2)	0.829	-13.759	15.416
	5 (2)	2.249	-12.338	16.836
	6 (2)	-3.078	-17.665	11.510

None of the random coefficients are significant, since the *p-values* are quite large (larger than 0.05, say). So it makes sense to look into the main effects. These were

tested in fixed effect table given earlier. It showed that the methods indeed differed significantly.

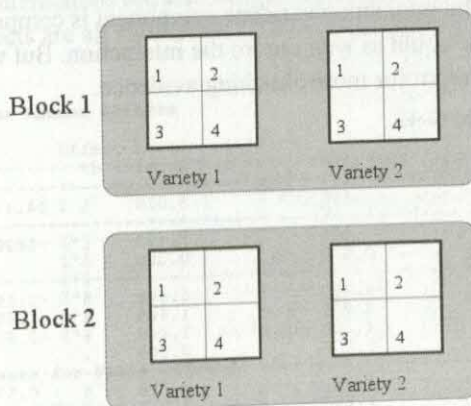
Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
METHOD	1	4	39.720	0.003

Once more we see that the methods produce significantly different measurements from the Type III Tests for the Fixed Effects table.

Example 4 Split Plot Design

The data set for this example comes from Milliken and Johnson (1992). It is an agricultural data set obtained from a split plot design laid out as follows. We want to compare four fertilizers and two varieties of crops. We have 4 (whole) plots to try these on. These are grouped into two blocks. The two varieties are assigned randomly to the two (whole) plots in each group. Each whole plot is split into 4 subplots, and the 4 fertilizers are applied randomly to these.



The yield of crop for each subplot is noted. The data are given in *CROPS* data file.

The input is:

```
USE CROPS
VC
CATEGORY BLOCK VARIETY FERT
MODEL YIELD = INTERCEPT + VARIETY + FERT + VARIETY * FERT
RANDOM BLOCK VARIETY * BLOCK
ESTIMATE
```

The following are selections from the output:

Fit Statistics

```
Final L-L      : -19.310
-2L-L         : 38.619
AIC           : 44.619
AIC(Corrected) : 50.619
BIC           : 44.857
```

Estimates of Covariance Components

Random Effect	Description	Estimate
BLOCK	Variance	16.087
	Parameter	
VARIETY*BLOCK	Variance	0.061
	Parameter	
Error variance	Variance	2.159

Notice how small the interaction variance component is compared with the other two. This signals that we could as well ignore the interaction. But we should wait for the latter parts of the output for more clinching evidence.

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		43.800	3.026	1	14.477	0.044
VARIETY	1	-1.800	1.490	1	-1.208	0.440
	2	0.000	0.000	.	.	.
FERT	1	-4.700	1.469	6	-3.199	0.019
	2	-3.900	1.469	6	-2.654	0.038
	3	-4.200	1.469	6	-2.858	0.029
	4	0.000	0.000	.	.	.
VARIETY*FERT	1*1	1.200	2.078	6	0.577	0.585
	1*2	1.600	2.078	6	0.770	0.471
	1*3	1.400	2.078	6	0.674	0.526
	1*4	0.000	0.000	.	.	.
	2*1	0.000	0.000	.	.	.
	2*2	0.000	0.000	.	.	.
	2*3	0.000	0.000	.	.	.
	2*4	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		43.800	36.397	51.203
VARIETY	1	-1.800	-5.446	1.846
	2	0.000	.	.
FERT	1	-4.700	-8.296	-1.104
	2	-3.900	-7.496	-0.304
	3	-4.200	-7.796	-0.604
	4	0.000	.	.
VARIETY*FERT	1*1	1.200	-3.885	6.285
	1*2	1.600	-3.485	6.685
	1*3	1.400	-3.685	6.485
	1*4	0.000	.	.
	2*1	0.000	.	.
	2*2	0.000	.	.
	2*3	0.000	.	.
	2*4	0.000	.	.

First let us understand why there are so many dots in this table. These are because of the estimability condition enforced by SYSTAT. The last coefficient of each fixed main effect is assumed to be 0. Also, many of the interaction terms have to be assumed 0 to keep the others estimable. The *p-values* are not reported for these forced zeros, since they would not make any sense there. The reported *p-values* for the interaction terms are all insignificant (larger than 0.05, say). As we shall presently see from a latter table, the random interactions are also insignificant. Let us now look at the main effects. The fertilizer effects are all significant (at 0.05 level) but the variety effects are not significant.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
BLOCK	1	-2.810	2.862	6	-0.982	0.364
	2	2.810	2.862	6	0.982	0.364
VARIETY*BLOCK	1*1	-0.045	0.243	6	-0.183	0.861
	1*2	0.045	0.243	6	0.183	0.861
	2*1	0.034	0.243	6	0.139	0.894
	2*2	-0.034	0.243	6	-0.139	0.894
	2*1	0.034	-0.562		0.630	
	2*2	-0.034	-0.630		0.562	

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
VARIETY	1	1	0.937	0.510
FERT	3	6	6.205	0.029
VARIETY*FERT	3	6	0.239	0.866

The ANOVA table usually provides a summary of what we have already found from the other tables. Here we see that the fertilizers differ significantly, while the other effects are insignificant.

Example 5 Using Covariates

This example is based on a clinical data set presented in Hocking (2003), where a pharmaceutical firm wants to test a new drug for a particular disease. The response is a measure of the improvement in the patient's status. A sample of 3 clinics is selected at random from a large population of clinics. From each clinic a sample of 10 patients with the particular disease are selected. The drug is applied to each patient and a response (Y) of the drug and a relevant physical characteristic (Z) for each patient, are recorded. The *CLINCOV* file contains this data set.

We want to fit the following model to this data set.

$$y_{ij} = \mu + \alpha_i + \beta z_{ij} + \varepsilon_{ij},$$

where α_i 's are the only random effects. The aim is to see if the drug is really effective or not, and whether the clinics influence the effect of the drug significantly. We want to guard against accidentally attributing any change in a patient's status to the drug if the change is actually due to the relevant physical characteristic of the patient. That is why we have included Z in our model.

The input is:

```
USE CLINCOV
VC
CATEGORY CLINIC
MODEL Y = INTERCEPT + Z
RANDOM CLINIC
ESTIMATE
```

The following are selections from the output:

Dimensions

```
Covariance Parameters : 3
Columns in X           : 2
Columns in Z           : 3
No. of Observations    : 30
```

The first column in X is due to the intercept term, while the second comes from the covariate.

Fit Statistics

Final L-L	: -75.717
-2L-L	: 151.434
AIC	: 155.434
AIC(Corrected)	: 155.914
BIC	: 158.099

Estimates of Covariance Components

Random Effect	Description	Estimate
CLINIC	Variance Parameter	0.001
Error variance	Variance Parameter	8.964

The first variance component is estimated to be 0 up to three decimal places. Compared to the much larger error variance this already foreshadows the insignificance of the clinic effect.

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	6.251	0.856	2	7.300	0.018
Z	0.567	0.083	26	6.803	0.000

Confidence Intervals of Fixed Effects Estimates

Effect	Estimate	95.00% Confidence Interval	
		Lower	Upper
Intercept	6.251	4.491	8.011
Z	0.567	0.396	0.739

Both the intercept and the slope terms are highly significant as judged by the small *p*-values. The significance of the intercept says that the drug has nontrivial effect, while the significant slope indicates that the patients' responses depend significantly on their physical characteristics as well.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
CLINIC	1	0.000	0.028	26	0.001	0.999
	2	0.000	0.028	26	0.009	0.993
	3	0.000	0.028	26	-0.010	0.992

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
CLINIC	1	0.000	-0.057	0.057
	2	0.000	-0.057	0.057
	3	0.000	-0.057	0.057

The clinic effects are all insignificant. So the clinics do not differ much among themselves in the present context.

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
Z	1	26	46.277	0.000

So far we have accepted the covariate as something informative. But is the covariate really bringing relevant information? The small *p-value* in the above F-test assures us this in the affirmative. Had this test shown the effect of Z to be insignificant we would have done better by dropping it from our model.

Example 6

Unbalanced Data: Different Types of ANOVA

This example is meant to show the difference among the three ANOVA methods: TYPE1, TYPE2 and TYPE3. These three methods will always give the same results if the data set is balanced, i.e, if there are equal number of observations in each cell. (For a nested design, a balanced data set also requires the same number of nested levels under each nesting effect.) So to see the difference among the three methods we need an unbalanced data set.

The *MACHINE2* data, from Milliken and Johnson (1992) p.285 presents an unbalanced data set where two machines are being operated by six randomly selected operators. Each operator is allowed to operate each machine at most three times.

Here machine is a fixed effect and operator is a random effect. In this unbalanced case the three methods can lead to different outcomes.

First we shall apply the TYPE1 method.

The input is:

```
USE MACHINE2
VC
CATEGORY MACHINE OPERATOR
MODEL SCORE = INTERCEPT + MACHINE
RANDOM OPERATOR + MACHINE*OPERATOR
ESTIMATE/ METHOD = TYPE1
```

Refer to chapter on "Linear Models" on page 1, Statistics II for a description of various types of sum of squares. We shall focus our attention on only the part of the output that

highlights the difference between the three types of ANOVA estimation. The rest of the output is not shown.

The following are selections from the output:

The categorical values encountered during processing are

Variables	Levels				
MACHINE (2 levels)	1.000	2.000			
OPERATOR (6 levels)	1.000	2.000	3.000	4.000	5.000
	6.000				

Dependent Variable : SCORE
 Fixed Factor(s) : MACHINE
 Fixed Covariate(s) : Intercept
 Random Factor(s) : OPERATOR, MACHINE*OPERATOR
 Estimation Method : ANOVA Type I

Dimensions

Covariance Parameters : 3
 Columns in X : 3
 Columns in Z : 18
 No. of Observations : 26

Type I Sum of Squares

Source	df	SS	Mean Squares	Expected MS
MACHINE	1	359.434	359.434	Q(MACHINE) + 0.136*V(OPERATOR) + 2.443*V(MA..)
OPERATOR	5	797.665	159.533	4.219*V(OPERATOR) + 2.176*V(MACHINE*OPERATOR..)
MACHINE*OPERATOR	5	288.030	57.606	2.043*V(MACHINE*OPERATOR) + V(Error)

* Q(): Quadratic term involving parameters of the fixed effects as indicated

This is a table that is produced whenever an ANOVA estimation method (of any type) is used. The first four columns are just as in any ANOVA table. The last column tells us what each MS is estimating unbiasedly. This information is important for making sense out of the ANOVA F-tests that are presented later. For unbalanced data the MS for each effect may have some contribution from some other effects mixed with it. This depends on the nature of imbalance and the type of ANOVA used. Here, for example,

the MS for machine has part of the operator effect and machine-operator interaction in its expectation.

Error Terms

Effect	Denominator Expression	Error Term
MACHINE	0.032 MS (OPERATOR) + 1.162 MS (MACHINE*OPE- R.)	71.875
OPERATOR	1.065 MS (MACHINE*OPE- RATOR) - 0.065 MS (Error)	61.302
MACHINE*OPERATOR	MS (Error)	0.880

Analysis of Variance

Source	Type I SS	Numerator df	Denominator df	Mean Squares	F-ratio
MACHINE	359.434	1	5.734	359.434	5.001
OPERATOR	797.665	5	4.991	159.533	2.602
MACHINE*OPERATOR	288.030	5	14.000	57.606	65.497
ERROR	12.313		14	0.880	

Analysis of Variance (contd...)

Source	p-value
MACHINE	0.069
OPERATOR	0.159
MACHINE*OPERATOR	0.000
ERROR	

Here is the ANOVA table. Only the interaction effect appears significant (*p-value* less than 0.05, say). Since we are using TYPE1 ANOVA, this means that interaction is significant after taking out the main effects of the machines and operators. The insignificant machine effect means that the machine effect is insignificant after taking out the operator effect and interaction effect. Thus, in a sense, the TYPE1 method is already taking the interaction into consideration before reporting the main effects.

Type I Tests for Fixed Effects

Source	Numerator df	Denominator df	F-ratio	p-value
MACHINE	1	5.734	6.369	0.047

The two tables above may appear contradictory at first. Both seem to test the machine effect, yet produce different *p-values*! The difference is explained by the fact that the first table performs the test with the random effects in mind. That is why the denominator degrees of freedom is fractional. However, the second table carries out simple fixed effect ANOVA.

Next let us use the TYPE2 method.

The input is:

```
USE MACHINE2
VC
CATEGORY MACHINE OPERATOR
MODEL SCORE = INTERCEPT + MACHINE
RANDOM OPERATOR + MACHINE*OPERATOR
ESTIMATE/ METHOD = TYPE2
```

The following are selections from the output:

```
Dependent Variable : SCORE
Fixed Factor(s) : MACHINE
Fixed Covariate(s) : Intercept
Random Factor(s) : OPERATOR, MACHINE*OPERATOR
Estimation Method : ANOVA Type II
```

Type II Sum of Squares

Source	df	SS	Mean Squares	Expected MS
MACHINE	1	284.311	284.311	$Q(\text{MACHINE}) + 2.318 \cdot V(\text{MACHINE*OPERATOR}) + V(\dots)$
OPERATOR	5	797.665	159.533	$4.219 \cdot V(\text{OPERATOR}) + 2.176 \cdot V(\text{MACHINE*OPERATOR}) + \dots$
MACHINE*OPERATOR	5	288.030	57.606	$2.043 \cdot V(\text{MACHINE*OPERATOR}) + V(\text{Error})$

* Q(): Quadratic term involving parameters of the fixed effects as indicated

This table shows the expected values of the ANOVA MS. A quick comparison with the corresponding table for TYPE 1 would show that the interaction MS and the operator MS are the same, but the machine MS is now different.

Error Terms

Effect	Denominator Expression	Error Term
MACHINE	1.135 $MS(\text{MACHINE*OPERATOR}) - 0.135$ $MS(\text{Error})$	65.255
OPERATOR	1.065 $MS(\text{MACHINE*OPERATOR}) - 0.065$ $MS(\text{Error})$	61.302
MACHINE*OPERATOR	$MS(\text{Error})$	0.880

Analysis of Variance

Source	Type II SS	Numerator df	Denominator df	Mean Squares	F-ratio
MACHINE	284.311	1	4.982	284.311	4.357
OPERATOR	797.665	5	4.991	159.533	2.602
MACHINE*OPERATOR	288.030	5	14.000	57.606	65.497
ERROR	12.313		14	0.880	

Analysis of Variance (contd...)

Source	p-value
MACHINE	0.091
OPERATOR	0.159
MACHINE*OPERATOR	0.000
ERROR	

Type II Tests for Fixed Effects

Source	Numerator df	Denominator df	F-ratio	p-value
MACHINE	1	4.982	6.369	0.053

Finally, here is what happens with the TYPE3 method.

The input is:

```
USE MACHINE2
VC
CATEGORY MACHINE OPERATOR
MODEL SCORE = INTERCEPT + MACHINE
RANDOM OPERATOR + MACHINE*OPERATOR
ESTIMATE/ METHOD = TYPE3
```

The following are selections from the output:

```
Dependent Variable : SCORE
Fixed Factor(s) : MACHINE
Fixed Covariate(s) : Intercept
Random Factor(s) : OPERATOR, MACHINE*OPERATOR
Estimation Method : ANOVA Type III
```

Variance Components Models

Dimensions

Covariance Parameters : 3
 Columns in X : 3
 Columns in Z : 18
 No. of Observations : 26

Type III Sum of Squares

Source	df	SS	Mean Squares	Expected MS
MACHINE	1	324.803	324.803	$Q(\text{MACHINE}) + 1.800 \cdot V(\text{MACHINE} \cdot \text{OPERATOR}) + V(\dots)$
OPERATOR	5	778.751	155.750	$4.086 \cdot V(\text{OPERATOR}) + 2.043 \cdot V(\text{MACHINE} \cdot \text{OPERATOR}) + \dots$
MACHINE*OPERATOR	5	288.030	57.606	$2.043 \cdot V(\text{MACHINE} \cdot \text{OPERATOR}) + V(\text{Error})$

* Q(): Quadratic term involving parameters of the fixed effects as indicated

Here the MS for both the main effects are different from those for types 1 and 2. However, the interaction MS is the same for all the three types. This is because, for all the three types the interaction SS is computed after taking out the contribution of the other two effects.

Error Terms

Effect	Denominator Expression	Error Term
MACHINE	0.881 $MS(\text{MACHINE} \cdot \text{OPERATOR}) + 0.119$ $MS(\text{Error})$	50.859
OPERATOR	1.000 $MS(\text{MACHINE} \cdot \text{OPERATOR}) + 0.000$ $MS(\text{Error})$	57.606
MACHINE*OPERATOR	$MS(\text{Error})$	0.880

Analysis of Variance

Source	Type III SS	Numerator df	Denominator df	Mean Squares	F-ratio
MACHINE	324.803	1	5.021	324.803	6.386
OPERATOR	778.751	5	5.000	155.750	2.704
MACHINE*OPERATOR	288.030	5	14.000	57.606	65.497
ERROR	12.313		14	0.880	

Analysis of Variance (contd...)

Source	p-value
--------	---------


```

MACHINE      : 0.053
OPERATOR     : 0.150
MACHINE*OPERATOR : 0.000
ERROR        :

```

Type III Tests for Fixed Effects

Source	Numerator df	Denominator df	F-ratio	p-value
MACHINE	1	5.021	6.369	0.053

Example 7

Exploring with Residuals

After drying beech wood the humidity level at any given point inside a plank typically depends on the depth of the point. In this example we want to study the relation between the humidity level (measured as a percentage) with the depth for 20 different randomly selected beech planks. For each plank we measure the humidity level for 5 depths and 3 widths. The *PLANKS* data file contains this data set.

We want to model the data as follows:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta + \varepsilon_{ijk},$$

where $i=1,2,3$, $j=1,\dots,5$ and $k=1,\dots,20$. Here the α 's denote the depth effect, the β 's denote the width effects and δ 's denote the plank effects. We have also allowed interaction between the depth and width effects. The interaction effect is denoted by the γ 's. We do not want our inference to involve the particular sample of 20 planks used in the experiment. So we consider the plank effect as random.

The input is:

```

USE PLANKS
VC
CATEGORY DEPTH WIDTH PLANK
MODEL HUMIDITY = INTERCEPT + WIDTH + DEPTH + WIDTH*DEPTH
RANDOM PLANK
ESTIMATE

```

The following are selections from the output:

Dimensions

```

Covariance Parameters : 3
Columns in X          : 24
Columns in Z          : 20
No. of Observations   : 300

```

The 24 columns in X have the following genesis: 1 column from the intercept term, 3 from the width coefficients, 5 from depth, and 3 times 5 from the interactions.

Fit Statistics

Final L-L : -332.029
 -2L-L : 664.058
 AIC : 668.058
 AIC (Corrected) : 668.100
 BIC : 675.363

Estimates of Covariance Components

Random Effect	Description	Estimate
PLANK	Variance Parameter	0.980
Error variance	Variance Parameter	0.404

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		4.395	0.263	19	16.710	0.000
WIDTH	1	0.300	0.201	266	1.493	0.137
	2	0.475	0.201	266	2.364	0.019
	3	0.000	0.000	.	.	.
DEPTH	1	0.130	0.201	266	0.647	0.518
	3	1.005	0.201	266	5.002	0.000
	5	1.315	0.201	266	6.544	0.000
	7	1.070	0.201	266	5.325	0.000
	9	0.000	0.000	.	.	.
WIDTH*DEPTH	1*1	-0.230	0.284	266	-0.809	0.419
	1*3	0.220	0.284	266	0.774	0.440
	1*5	0.325	0.284	266	1.144	0.254
	1*7	0.260	0.284	266	0.915	0.361
	1*9	0.000	0.000	.	.	.
	2*1	0.025	0.284	266	0.088	0.930
	2*3	0.520	0.284	266	1.830	0.068
	2*5	0.355	0.284	266	1.249	0.213
	2*7	0.160	0.284	266	0.563	0.574
	2*9	0.000	0.000	.	.	.
	3*1	0.000	0.000	.	.	.
	3*3	0.000	0.000	.	.	.
	3*5	0.000	0.000	.	.	.
	3*7	0.000	0.000	.	.	.
	3*9	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		4.395	3.877	4.913
WIDTH	1	0.300	-0.096	0.696
	2	0.475	0.079	0.871
	3	0.000	.	.
DEPTH	1	0.130	-0.266	0.526
	3	1.005	0.609	1.401
	5	1.315	0.919	1.711
	7	1.070	0.674	1.466
	9	0.000	.	.

WIDTH*DEPTH	1*1	-0.230	-0.789	0.329
	1*3	0.220	-0.339	0.779
	1*5	0.325	-0.234	0.884
	1*7	0.260	-0.299	0.819
	1*9	0.000	.	.
	2*1	0.025	-0.534	0.584
	2*3	0.520	-0.039	1.079
	2*5	0.355	-0.204	0.914
	2*7	0.160	-0.399	0.719
	2*9	0.000	.	.
	3*1	0.000	.	.
	3*3	0.000	.	.
	3*5	0.000	.	.
	3*7	0.000	.	.
	3*9	0.000	.	.

This table shows, among other things, the results of the t-tests for each fixed effect coefficient. As usual we start by considering the higher-order effects first. The interactions are all insignificant (*p-values* above 0.05, say). The dots in some of the rows owe their origin to the estimability restriction imposed by SYSTAT. Next we look at the main effects. Width2 appears to be only significant width term. The fact that the *p-value* for Width1 is large means Width1 does not differ significantly from Width3, which is the reference Width (since SYSTAT assumes that the coefficient for Width3 to be 0 as an estimability restriction). A similar argument shows that coefficients of Depths1 and 9 do not differ significantly, while the other depth coefficients do.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
PLANK	1	-0.882	0.272	266	-3.245	0.001
	2	-0.597	0.272	266	-2.195	0.029
	3	-0.194	0.272	266	-0.715	0.475
	4	0.208	0.272	266	0.765	0.445
	5	-1.012	0.272	266	-3.723	0.000
	6	0.279	0.272	266	1.028	0.305
	7	-0.091	0.272	266	-0.333	0.739
	8	-0.837	0.272	266	-3.078	0.002
	9	-1.304	0.272	266	-4.797	0.000
	10	1.850	0.272	266	6.805	0.000
	11	-0.486	0.272	266	-1.789	0.075
	12	0.124	0.272	266	0.455	0.650
	13	0.902	0.272	266	3.319	0.001
	14	0.467	0.272	266	1.720	0.087
	15	1.882	0.272	266	6.924	0.000
	16	0.714	0.272	266	2.627	0.009
	17	-1.096	0.272	266	-4.033	0.000
	18	1.045	0.272	266	3.845	0.000
	19	-1.453	0.272	266	-5.346	0.000
	20	0.480	0.272	266	1.768	0.078

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
PLANK	1	-0.882	-1.417	-0.347
	2	-0.597	-1.132	-0.061
	3	-0.194	-0.729	0.341
	4	0.208	-0.327	0.743
	5	-1.012	-1.547	-0.477

6	0.279	-0.256	0.814
7	-0.091	-0.626	0.445
8	-0.837	-1.372	-0.302
9	-1.304	-1.839	-0.769
10	1.850	1.314	2.385
11	-0.486	-1.021	0.049
12	0.124	-0.412	0.659
13	0.902	0.367	1.437
14	0.467	-0.068	1.003
15	1.882	1.347	2.417
16	0.714	0.179	1.249
17	-1.096	-1.631	-0.561
18	1.045	0.510	1.580
19	-1.453	-1.988	-0.918
20	0.480	-0.055	1.016

Most of the plank effect coefficients are significant (p -values below 0.05, say).

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
WIDTH	2	266	29.646	0.000
DEPTH	4	266	78.259	0.000
WIDTH*DEPTH	8	266	1.084	0.375

The interaction effect is not significant. But the other two main fixed effects are significant. So we drop the interaction from the model and refit.

The input is:

```
VC
CATEGORY DEPTH WIDTH PLANK
MODEL HUMIDITY = INTERCEPT + WIDTH + DEPTH
RANDOM PLANK
SAVE MYRESIDS / MRESIDUALS
ESTIMATE
```

Notice that here we are saving the marginal residuals in a data set called *MYRESIDS*, which will be automatically created by SYSTAT. This data set will have two variables: estimate and mresiduals.

The following are selections from the output:

Fit Statistics

```
Final L-L      : -331.912
-2L-L          : 663.824
AIC            : 667.824
AIC(Corrected) : 667.865
BIC            : 675.184
```


Estimates of Covariance Components

Random Effect	Description	Estimate
PLANK	Variance Parameter	0.980
Error variance	Variance Parameter	0.405

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		4.286	0.242	19	17.731	0.000
WIDTH	1	0.415	0.090	274	4.613	0.000
	2	0.687	0.090	274	7.636	0.000
	3	0.000	0.000	.	.	.
DEPTH	1	0.062	0.116	274	0.531	0.596
	3	1.252	0.116	274	10.776	0.000
	5	1.542	0.116	274	13.273	0.000
	7	1.210	0.116	274	10.417	0.000
	9	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		4.286	3.810	4.762
WIDTH	1	0.415	0.238	0.592
	2	0.687	0.510	0.864
	3	0.000	.	.
DEPTH	1	0.062	-0.167	0.290
	3	1.252	1.023	1.480
	5	1.542	1.313	1.770
	7	1.210	0.981	1.439
	9	0.000	.	.

This table shows, among other things, the results of the t-tests for each fixed effect coefficient. Notice that dropping the interaction term has changed things considerably. Coefficients for Widths1 and 3 are now significantly different. However, the coefficients of Depths1 and 9 still do not differ significantly, while the other depth coefficients do.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
PLANK	1	-0.882	0.272	274	-3.244	0.001
	2	-0.597	0.272	274	-2.194	0.029
	3	-0.194	0.272	274	-0.715	0.475
	4	0.208	0.272	274	0.765	0.445
	5	-1.012	0.272	274	-3.721	0.000
	6	0.279	0.272	274	1.027	0.305
	7	-0.091	0.272	274	-0.333	0.739
	8	-0.837	0.272	274	-3.077	0.002
	9	-1.304	0.272	274	-4.795	0.000
	10	1.849	0.272	274	6.802	0.000
	11	-0.486	0.272	274	-1.788	0.075
	12	0.124	0.272	274	0.455	0.650
	13	0.902	0.272	274	3.318	0.001

14	0.467	0.272	274	1.719	0.087
15	1.882	0.272	274	6.921	0.000
16	0.714	0.272	274	2.626	0.009
17	-1.096	0.272	274	-4.031	0.000
18	1.045	0.272	274	3.843	0.000
19	-1.453	0.272	274	-5.344	0.000
20	0.480	0.272	274	1.767	0.078

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
PLANK	1	-0.882	-1.417	-0.347
	2	-0.597	-1.132	-0.061
	3	-0.194	-0.730	0.341
	4	0.208	-0.327	0.743
	5	-1.012	-1.547	-0.477
	6	0.279	-0.256	0.815
	7	-0.091	-0.626	0.445
	8	-0.837	-1.372	-0.301
	9	-1.304	-1.839	-0.768
	10	1.849	1.314	2.385
	11	-0.486	-1.022	0.049
	12	0.124	-0.412	0.659
	13	0.902	0.367	1.437
	14	0.467	-0.068	1.003
	15	1.882	1.347	2.417
	16	0.714	0.179	1.249
	17	-1.096	-1.631	-0.561
	18	1.045	0.510	1.580
	19	-1.453	-1.988	-0.918
	20	0.480	-0.055	1.016

Most of the plank effect coefficients are significant (p-values below 0.05, say).

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
WIDTH	2	274	29.574	0.000
DEPTH	4	274	78.068	0.000

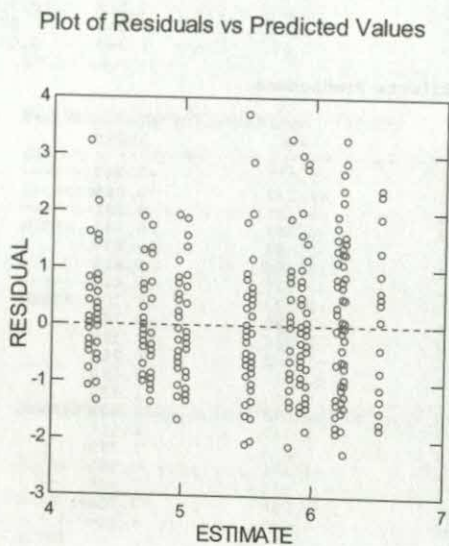
Next we shall look at the residual plot, where residuals are plotted against estimates. If the Quick Graph feature is turned on then the plot is automatically produced. Otherwise you may create the plot by an explicit PLOT command as shown below. Incidentally, if you are using the Quick Graph feature then you do not need to save the residuals as we did above.

To make the plot directly,

The input is:

```
USE MYRESIDS
PLOT MRESIDUAL * ESTIMATE
```

The output is:



Example 8 Missing Data

This example shows how we can deal with missing values in a data set using SYSTAT. The data set we shall use is from Hocking (2003), where a pharmaceutical company is trying to test a new medicine. Three clinics have been selected at random from a large number of clinics. The drug is administered to 10 randomly selected patients. However, some of the measurements from some of the clinics have not been reported. The data are setup in the file *PATMISS*.

Before we can fit a statistical model to this data set, we need to have an idea about why some of the observations are missing. For instance, it may be because of some transcription problem which may be assumed to be independent of the observations. This called *Missing Completely At Random (MCAR)*, and is the most frequently made assumption in case we are ignorant about the exact cause behind the data loss. SYSTAT, like most other software, analyzes this special case only. If, however, the missing observations correspond to patients for whom the drug have caused serious side-effects leading to cancelling the medication, then the analyst had better investigate

the nature of the side effects, rather than continue happily with the MCAR assumption about the incomplete data set.

In this data set, however, no such cause is reported against a reasonable use of the MCAR assumption.

The input is:

```
USE PATMISS
VC
  CATEGORY CLINIC
  MODEL Y = INTERCEPT + Z
  RANDOM CLINIC
  ESTIMATE
```

Notice that no mention is made about the data being incomplete. This is because, SYSTAT sees the incomplete nature of the data set from the data set itself, and then it automatically uses the appropriate analysis using the MCAR assumption.

The output is:

Categorical values encountered during processing are

Variables	Levels
CLINIC (3 levels)	1.000 2.000 3.000

```
Dependent Variable : Y
Fixed Covariate(s) : Intercept
Random Factor(s)   : CLINIC
Estimation Method  : Residual or Restricted Maximum Likelihood (REML)
```

Dimensions

```
Covariance Parameters : 2
Columns in X           : 1
Columns in Z           : 3
No. of Observations   : 20
```

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		124.120	
1	ECME	123.725	0.003
2	ECME	123.537	0.002
3	ECME	123.443	0.001
4	ECME	123.392	0.000
5	ECME	123.362	0.000
6	NR	123.314	0.001
7	NR	123.311	0.000
8	NR	123.311	0.000
9	NR	123.311	0.000

Fit Statistics

Final L-L : -61.655
 -2L-L : 123.311
 AIC : 127.311
 AIC(Corrected) : 128.061
 BIC : 129.200

Estimates of Covariance Components

Random Effect	Description	Estimate
CLINIC	Variance	6.265
	Parameter	
Error variance	Variance	30.222
	Parameter	

Estimates of Fixed Effects

Effect	Estimate	Standard Error	df	t	p-value
Intercept	6.762	1.930	2	3.504	0.073

Confidence Intervals of Fixed Effects Estimates

Effect	Estimate	95.00% Confidence Interval	
		Lower	Upper
Intercept	6.762	-1.541	15.066

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
CLINIC	1	-0.986	1.932	17	-0.510	0.616
	2	-1.162	1.984	17	-0.585	0.566
	3	2.148	2.047	17	1.049	0.309

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
CLINIC	1	-0.986	-5.063	3.090
	2	-1.162	-5.348	3.025
	3	2.148	-2.171	6.467

This example shows that the SYSTAT output for missing data is as simple as that for complete data. We see that the clinics do not differ significantly among themselves.

References

- Hocking, R. R. (2003). *Methods and applications of linear models*. New York: John Wiley & Sons.
- Kuehl, R. O. (2000). *Design of Experiments: Statistical principles of research design and analysis*. New York: Duxbury Thomson Learning.
- Milliken, G. A. and Johnson, D. E. (1992). *Analysis of messy data, Volume I: Designed Experiments*. London: Chapman & Hall.

Linear Mixed Models

Arnab Chakraborty, Ravindra Jore, Bindu-Madhav Yeelarathi, and Javed Pathan

Linear Mixed Models (LMM) fits and analyzes mixed models with structured covariance/correlation matrices for random effects and residuals. Variance Component, Compound Symmetry, Diagonal, and Unstructured are the four types of structures provided for random effects. Variance Components, Compound Symmetry, and Auto-Regressive(1) are the three types of structures provided for error covariances. Various models like random intercept model, random coefficients model, variance components model, mixed effects ANOVA model, and models with autocorrelated errors can be fitted using LMM. LMM allows random effects to be both categorical and continuous. In LMM, SYSTAT provides two methods to estimate covariance parameters, viz., Maximum Likelihood and Restricted (Residual) Maximum likelihood. SYSTAT provides:

- Covariance parameter estimates
- Fixed effect and random effect predictors, standard errors, confidence intervals and t-test for testing whether these estimates/predictors are significant.
- F-ratio tests for fixed effects.
- Log-likelihood, Akaike Information Criterion (AIC), Akaike Information Criterion Corrected (AICc) and Bayesian Information Criterion (BIC), and iteration history as default output. Save option is provided to save residuals, predictions, model parameter estimates and their standard errors to a specified data file.
- Plot of residuals against estimates

Statistical Background

A general linear mixed model is a model of the form

$$y = X\beta + Z_1\gamma_1 + \dots + Z_p\gamma_p + \varepsilon$$

where y is a response vector, X and Z_i 's are known matrices (either design matrices or covariate matrices), β is the vector of fixed effects, each γ_i is a vector of random effects, and ε is the random error vector. Here y is a random vector, whose randomness comes partly from the random vector γ_i and partly from ε . We assume that the random vectors γ_i and ε have independent Gaussian distributions with zero mean and covariance matrices having some user-specified structure. Each γ_i consists of the random coefficients for i th random effect. The variance-covariance matrix structure may be different for the different effects. SYSTAT provides the option to specify a common covariance parameter for multiple effects. MIXED offers two general estimation techniques: ML and REML. Both these methods are iterative. The latter produces unbiased estimators. For details of these methods, please refer to Chapter 5: Mixed models: Introduction in this volume.

SYSTAT reports the BLUES of the fixed effects and BLUPs of the random effects, as well as estimates of the variance parameters.

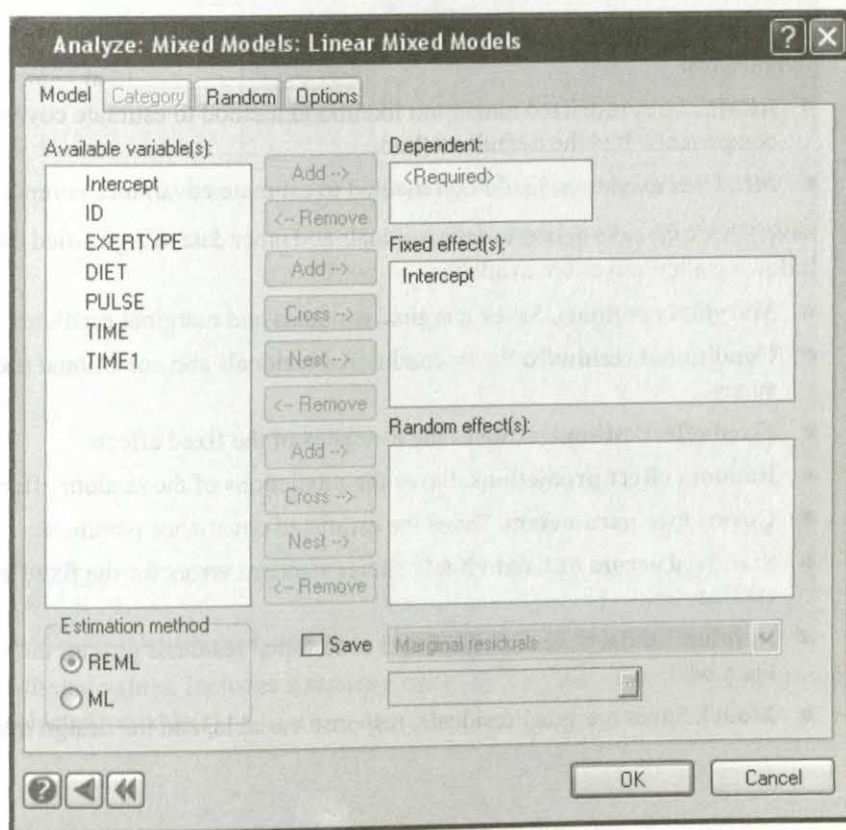
For each model you fit using MIXED, SYSTAT reports log-likelihood, Akaike Information Criterion (AIC), Bayes Information Criterion (BIC), and Akaike Information Criterion corrected (AICc) for assessing the fit of the model.

Linear Mixed Models in SYSTAT

Model Estimation (in MIXED)

To specify a linear mixed model using MIXED, from the menus choose:

Analyze
Mixed Models
Linear Mixed Models...



Dependent. Dependent is the variable you want to examine. Dependent variable should be a continuous numeric variable.

Fixed effect(s). Select one or more continuous or categorical (grouping) variables which you treat as fixed effects. Fixed effects that are not denoted as categorical are considered covariates. By default fixed intercept is present in model. If you want crossed or nested effects in your model, you need to build these components using Cross and Nest buttons.

Random effect(s). Select one or more continuous or categorical (grouping) variables which you treat as random effects. Random effects that are not denoted as categorical are considered covariates. If you want interactions or nested effects in your model, you need to build these components using Cross and Nest buttons.

Estimation method. Choose one among the available methods to estimate variance components.

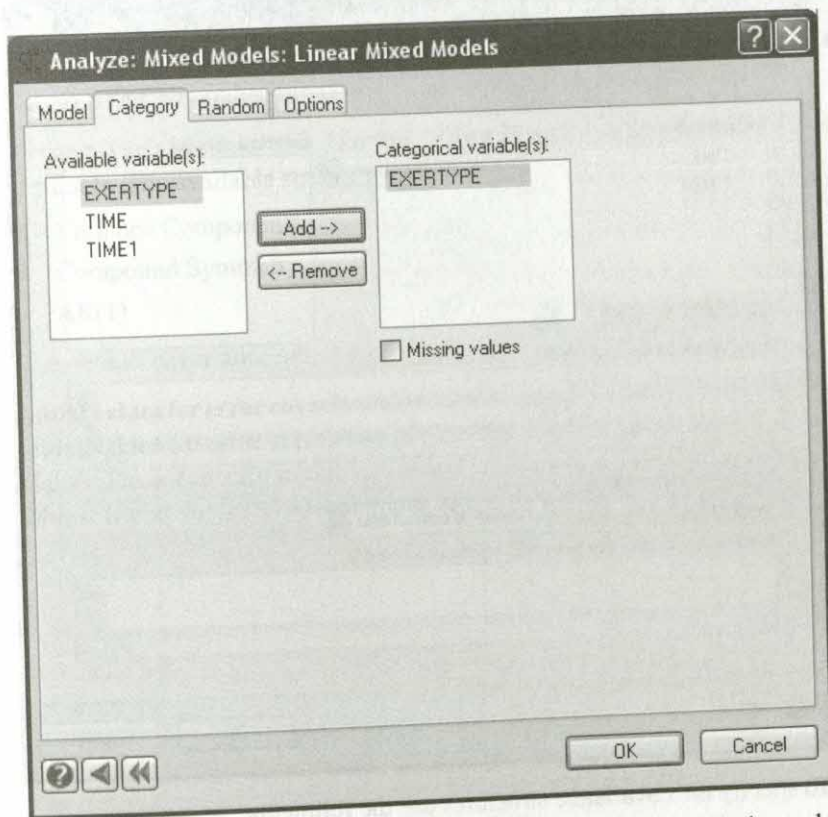
- **REML.** Uses restricted maximum likelihood method to estimate covariance components. It is the default method.
- **ML.** Uses maximum likelihood method to estimate covariance components.

Save. Check the save option to save residuals and other data to a specified data file. The following alternatives are available:

- **Marginal residuals.** Saves marginal residuals and marginal predicted values.
- **Conditional residuals.** Saves conditional residuals and conditional predicted values.
- **Fixed effect estimates.** Saves the estimates of the fixed effects.
- **Random effect predictions.** Saves the predictions of the random effects.
- **Covariance parameters.** Saves the estimated covariance parameters.
- **Standard errors of fixed effects.** Saves standard errors for the fixed effect estimates.
- **Residuals/data.** Saves marginal and conditional residuals plus all the variables in the model.
- **Model.** Saves marginal residuals, response variable, and the design matrices.

Category

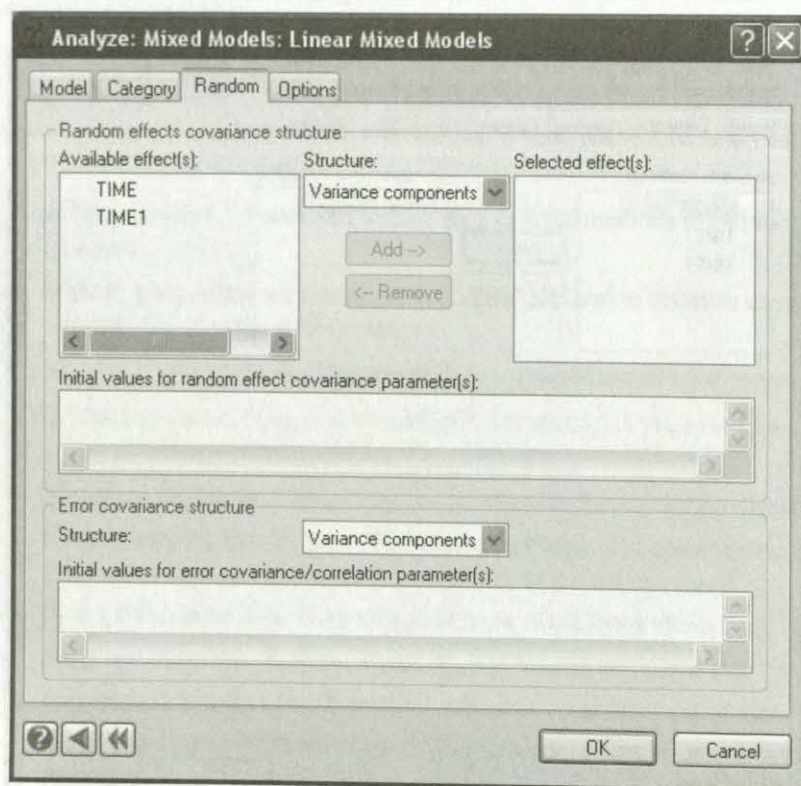
To specify categorical variables, click the **Category** tab. Select at least one fixed or random effect in **Model** tab other than intercept to activate this tab.



Missing values. Includes a separate category for cases with a missing value for the selected variable(s).

Random

To specify covariance structures for random effects and errors, click the Random tab.



To specify the covariance structures use the following:

Available effect(s). Available effects are the variables selected as random effects.

Structure. Choose a random effect and select one of the covariance structures available. The available structures are as follows:

- Variance Components
- Diagonal
- Compound Symmetry
- Unstructured

The default covariance structure is Variance Components.

Selected effect(s). Effect or effects along with their covariance structures.

Initial values for random effect covariance parameter(s). Use this option to provide initial values for covariance parameters. Specify values for each component in the order the effects appear in your model. Separate the values with commas or blanks. Do not specify initial values for some of the parameters and leave blanks for others. If you do, SYSTAT computes initial values for all covariance components.

Error covariance structure. Use this option to provide a covariance structure for residuals. The available structures are as follows:

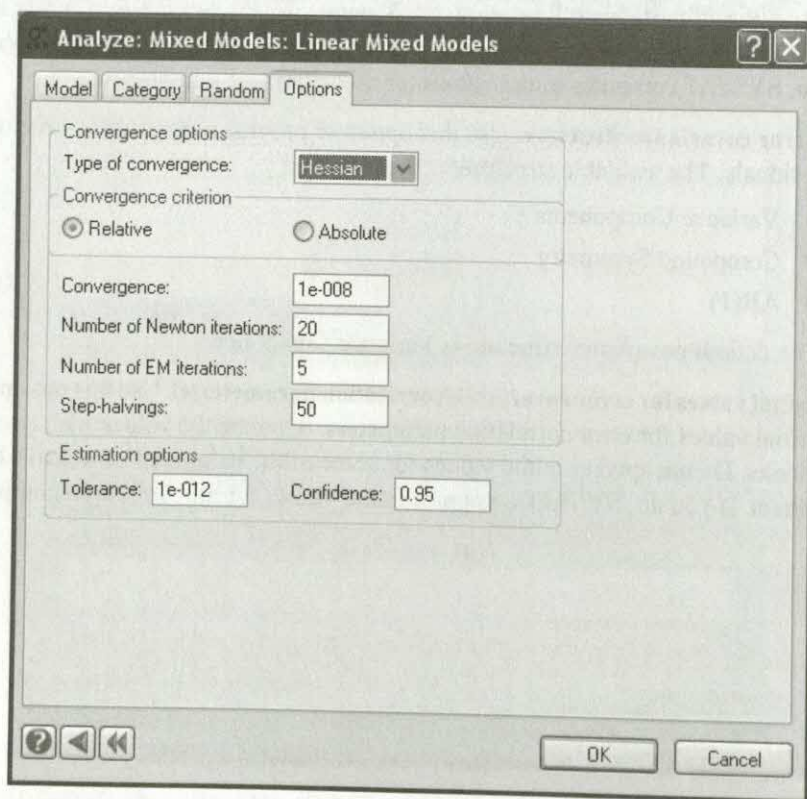
- Variance Components
- Compound Symmetry
- AR(1)

The default covariance structure is Variance Components.

Initial values for error covariance/correlation parameter(s). Use this option to provide initial values for error correlation parameters. Separate the values with commas or blanks. Do not specify initial values for some of the parameters and leave blanks for others. If you do, SYSTAT computes initial values for all correlation components.

Options

Use the Options tab to specify computational controls for REML or ML method of estimation.



SYSTAT offers the following options for controlling estimation using ML/REML:

Type of convergence. Check one of the following options to check convergence.

Three types of convergence checks are available:

- **Hessian.** Uses a quadratic form $g' H^{-1} g$ where g is the gradient vector and H is the hessian matrix.
- **Likelihood.** Uses the difference between log-likelihood at current iteration and the log-likelihood at last iteration.

- **Parameter.** Uses maximum of absolute differences between parameter estimates at current iteration and parameter estimates at last iteration.

Convergence criterion. Two criteria are available:

- **Relative.** Checks relative difference for convergence. That is, convergence checking is done relative to log-likelihood. It is the default option.
- **Absolute.** Tests convergence directly against a value specified.

Convergence. Specify a positive number. SYSTAT stops iterations when convergence value is less than this number.

Number of Newton iterations. Use this to specify maximum number of Newton-Raphson iterations for fitting your model. The default is 20.

Number of EM iterations. Use this to specify maximum number of EM iterations before going to Newton-Raphson iterations. Sufficient number of EM iterations provide good initial estimates for Newton-Raphson iterations. The default is 5.

Step-halvings. Use this to specify maximum number of step halvings. The default is 50.

Tolerance. A check for near singularity. Use Tolerance to guard against singularity problems.

Confidence. Specifies the confidence coefficient for testing purposes. The default is 0.95.

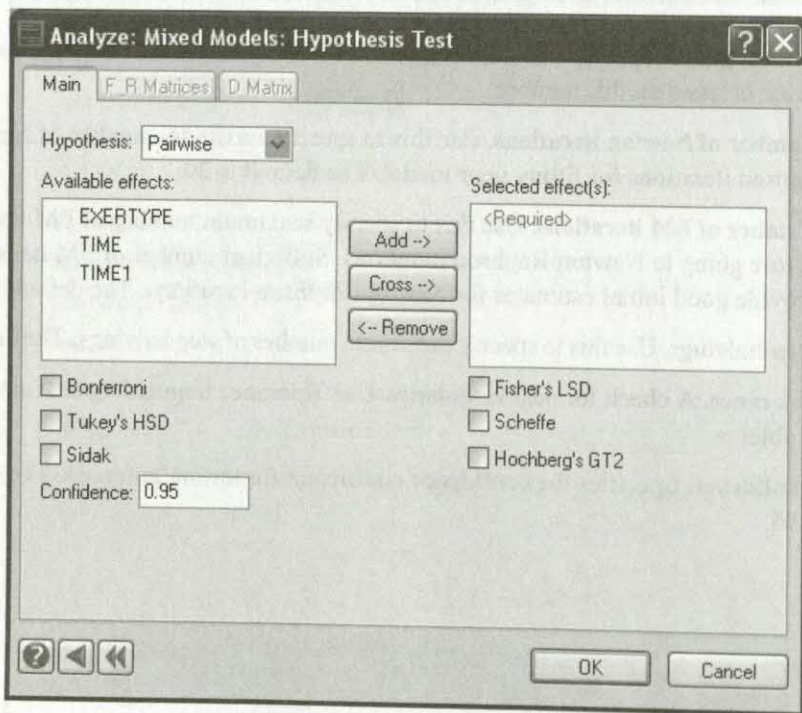
Hypothesis Tests

To test hypotheses, from the menus choose:

Analyze

Mixed Models

Hypothesis Test...



You can customize the hypothesis to be tested. Contrasts can be defined across the categories of a grouping factor.

Hypothesis. Select the type of hypothesis. The following choices are available:

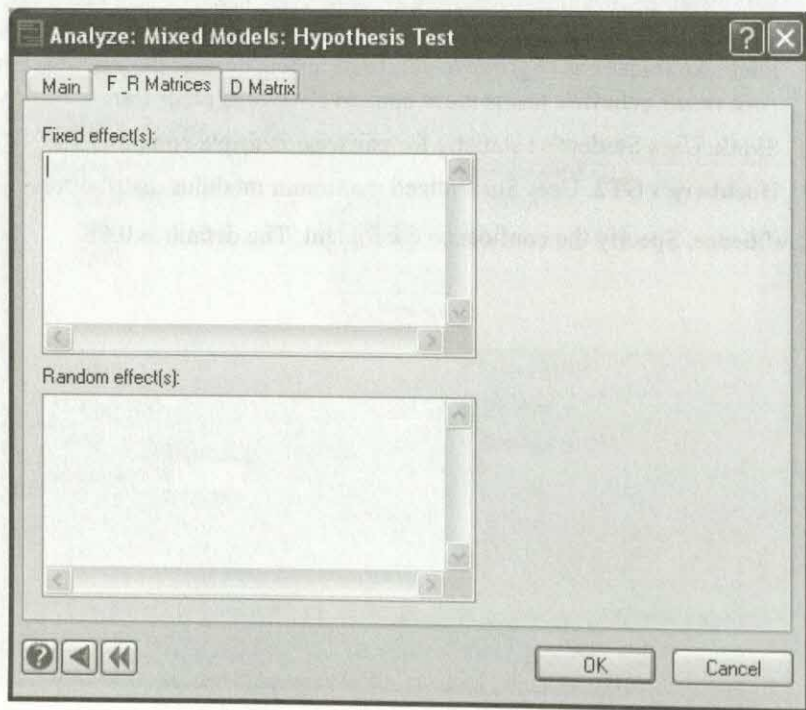
- **Pairwise.** Compare pairs of groups to determine which pairs differ.
- **F and R Matrices.** Tests the hypotheses corresponding to the F and R Matrices tab.

The following options are available to compute p-values adjusted for multiple comparisons:

- **Bonferroni.** Uses Student's t statistic. It sets the family-wise error rate as $(1 - \text{Confidence}) / (\text{Total number of comparisons})$.
 - **Tukey's HSD.** Uses the Studentized range statistic to make all pairwise comparisons. This is the default method.
 - **Fisher's LSD.** Equivalent to multiple t tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.
 - **Scheffé.** The significance level of Scheffé's test is designed to allow all possible linear combinations of group means to be tested, not just the pairwise comparisons. As a result Scheffé's test is more conservative than other tests.
 - **Sidak.** Uses Student's t statistic for pairwise multiple comparisons.
 - **Hochberg's GT2.** Uses Studentized maximum modulus distribution.
- Confidence.** Specify the confidence coefficient. The default is 0.95.

F and R Matrices

To specify **F** and **R** matrices, select the **F and R matrices** option of Hypothesis in the Mixed Models: Hypothesis Test dialog box. **F** and **R** are the matrices of linear weights contrasting the coefficient estimates for fixed and random effects respectively. You can write your hypothesis in terms of the **F** and **R** matrices.

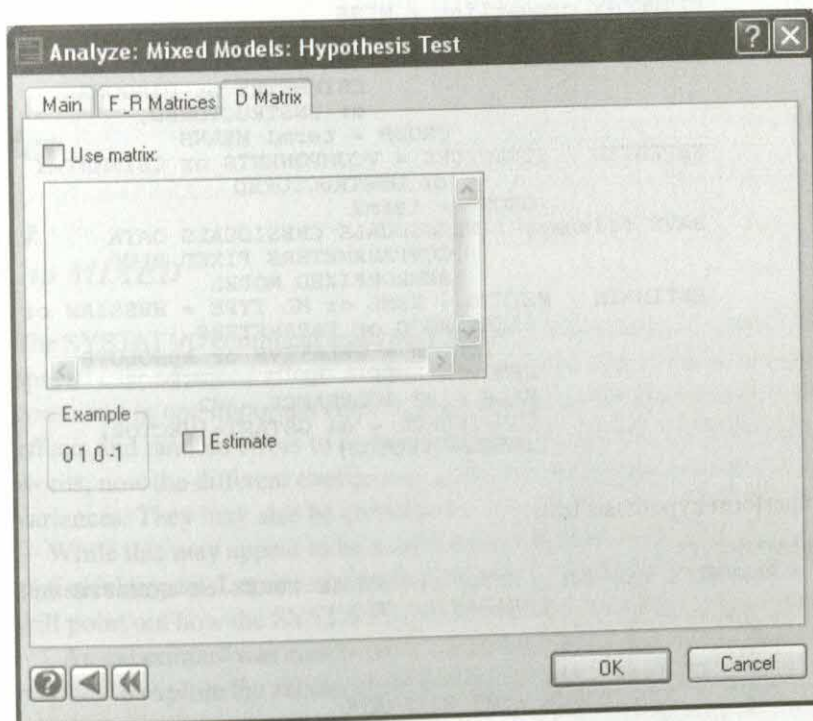


Fixed effect(s). Specify as many numbers as the dimension of your beta vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.

Random effect(s). Specify as many numbers as dimension of your gamma vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.

D Matrix

D is a null hypothesis vector. By default it is null vector. The **D** vector, if you use it, must have the same number of rows as the **F** or **R** matrices. To specify a different **D** Matrix, click the D Matrix tab in the Mixed Models: Hypothesis Test dialog box.



Specify a vector of dimension same as the number of rows in **F** and **R** matrices.

Estimate. Check this option for testing significance of contrasts (rows) in **F** and **R** matrices individually. This test reports estimate of the estimable linear parametric function, its standard error and the corresponding t-test.

Using Commands

To analyze linear mixed models using commands first select a data set with USE filename and continue with:

```
MIXED
  RESET
  CATEGORY grpvarlist / MISS
  MODEL var = INTERCEPT + varlist1
  RANDOM varlist2 / STRUCTURE = VCOMPONENTS or
                           CSYMMETRY or DIAGONAL
                           or UNSTRUCTURED,
                           GROUP = term1 MEANS
  REPEATED / STRUCTURE = VCOMPONENTS or CSYMMETRY
                           or UNSTRUCTURED
                           GROUP = term2
  SAVE filename / MRESIDUALS CRESIDUALS DATA
                  COVPARAMETERS FIXED BLUP
                  ERRORFIXED MODEL
  ESTIMATE / METHOD = REML or ML TYPE = HESSIAN or
              LIKELIHOOD or PARAMETERS
              CRITERION = RELATIVE or ABSOLUTE
              NEM = n1 NNR = n2 CONVERGENCE = d1
              HALF = n3 TOLERANCE = d2
              CONFIDENCE = d4 GSTART=[VECTOR]
              RSTART=[VECTOR]
```

To perform hypothesis tests:

```
HYPOTHESIS
  PAIRWISE varlist / BONF or LSD or TUKEY or SCHEFFE or
                  SIDAK or GT2
  FMATRIX [matrix]
  RMATRIX [matrix]
  DMATRIX [matrix]
  TEST / CONFIDENCE = d5 ESTIMATE
```

Usage Considerations

Types of data. MIXED requires a rectangular data file.

Print options. MIXED displays covariance parameters and tests of fixed effects for PLENGTH SHORT. For PLENGTH MEDIUM, MIXED adds fixed effects estimates. For PLENGTH LONG, MIXED adds random effects predictions and iteration history.

Quick Graphs. MIXED produces a Quick Graph of marginal residuals versus marginal predicted values.

Saving files. Several sets of output can be saved to a file. The actual contents of the saved file depend on the analysis. Files may include estimated regression coefficients, model variables, residuals, and predicted values.

BY groups. MIXED analyzes data by groups.

Case frequencies. MIXED uses the FREQUENCY variable, if present, to duplicate cases.

Case weights. MIXED uses the values of any WEIGHT variables to weight each case.

Examples

Example 1

From VC to MIXED

The SYSTAT VC command analyzes variance components models, which constitute a special case of mixed effects models. The MIXED command generalizes the VC command in one important respect; it allows the covariance matrices of the random effects and random errors to be other than multiples of the identity matrix. In other words, now the different coefficients under the same random effect can have different variances. They may also be correlated.

While this may appear to be a mild generalization mathematically, it has a deep statistical impact. Let us consider the following data set to illustrate this. This example will point out how the SYSTAT output for MIXED differs from that of VC.

An experiment was conducted by students at The Ohio State University in the fall of 1993 to explore the relationship between a person's heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable, heart rate, was measured in beats per minute. There were two different step heights: 5.75 inches (coded as 0), and 11.5 inches (coded as 1). There were three rates of stepping: 14 steps/min. (coded as 0), 21 steps/min. (coded as 1), and 28 steps/min. (coded as 2). This resulted in six possible height/frequency combinations. Each subject performed the activity for three minutes. Subjects were kept on pace by the beat of an electric metronome. One experimenter counted the subject's pulse for 20 seconds before and after each trial. The subject always rested between trials until their heart rate returned to close to the beginning rate. Another experimenter kept track of the time spent stepping. Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment. Each pair of experimenters was

treated as a block. The data are stored in the SYSTAT data file named *HEART*. The source of the data is the website CMU:DASL (2005).

Consider the model

$$y_{ijkl} = \mu + \alpha_{ij} + \beta_k + \varepsilon_{ijkl}$$

where y_{ijkl} is the l -th observation in the k -th block for the i -th height and j -th frequency. Here α_{ij} is the combined effect of the i -th height and j -th frequency, and β_k is the effect of the k -th block. We shall treat the α_{ij} 's as fixed.

Let us see if we should consider BLOCK as a random effect here. In VC, we can do so if we consider the experimenters as a random sample from an infinite population of experimenters. In other words, in future fresh replications of the experiment, we may use other experimenters. In that scenario, however, we must treat the β_k 's as independently and identically distributed random variables. This necessarily leads to a covariance matrix of the form $\sigma^2 I$, the form that characterizes variance components models. Under what situation then can we treat BLOCK as a random effect, and yet have some other covariance structure?

The answer to this question is the key to understanding when to use the MIXED command. Suppose that we do not have any more experimenters at hand, and so all future replications must use the same set of experimenters. However, even then we should consider the BLOCK effect as random, if we want our model to account for the fact that, for a different replication of the same experiment, the same experimenter may behave in a slightly different way (say, depending on their mood, which is random.) In this case, β_1 is a random variable for the first block, while β_2 is a random variable for the second block. In this situation the β 's may have different variances, even though we may still consider them as independent. The VC command is unable to tackle this model. So we shall employ the more powerful MIXED command here.

The input is:

```
USE HEART
LET Y=HR-RESTR
MIXED
CATEGORY HEIGHT FREQUENCY BLOCK
MODEL Y = INTERCEPT + HEIGHT*FREQUENCY
RANDOM BLOCK / STRUCTURE = DIAG
ESTIMATE
```

The output is:

Dependent Variable : Y
 Fixed Factor(s) : HEIGHT*FREQUENCY
 Fixed Covariate(s) : Intercept
 Random Factor(s) : BLOCK
 Estimation Method : Residual or Restricted Maximum Likelihood (REML)

Dimensions

Covariance Parameters : 7
 Columns in X : 7
 Columns in Z : 6
 No. of Observations : 30

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		183.775	
1	NR	182.972	0.006
2	NR	181.989	0.010
3	NR	181.652	0.003
4	NR	181.176	0.005
5	NR	180.940	0.002
6	NR	180.810	0.001
7	NR	180.740	0.001
8	NR	180.701	0.000
9	NR	180.680	0.000
10	NR	180.669	0.000
11	NR	180.663	0.000
12	NR	180.660	0.000
13	NR	180.658	0.000
14	NR	180.657	0.000
15	NR	180.657	0.000
16	NR	180.657	0.000
17	NR	180.657	0.000

Fit Statistics

Final L-L : -90.328
 -2L-L : 180.657
 AIC : 194.657
 AIC (Corrected) : 201.657
 BIC : 202.903

Estimates of Covariance Components

Random Effect	Description	Estimate
BLOCK	Variance 1	128.136
	Variance 2	0.005
	Variance 3	60.924
	Variance 4	3.794
	Variance 5	0.005
	Variance 6	303.873
Error variance	Variance	53.599
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		59.231	3.503	5	16.908	0.000
HEIGHT*FREQUENCY	0*0	-46.056	4.715	19	-9.768	0.000
	0*1	-33.688	4.694	19	-7.177	0.000
	0*2	-24.062	4.720	19	-5.097	0.000
	1*0	-29.058	4.688	19	-6.199	0.000
	1*1	-24.258	4.688	19	-5.175	0.000
	1*2	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		59.231	51.899	66.564
HEIGHT*FREQUENCY	0*0	-46.056	-55.924	-36.187
	0*1	-33.688	-43.511	-23.864
	0*2	-24.062	-33.941	-14.182
	1*0	-29.058	-38.870	-19.246
	1*1	-24.258	-34.069	-14.446
	1*2	0.000	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
BLOCK	1	-10.711	3.661	19	-2.926	0.009
	2	0.000	0.070	19	0.002	0.998
	3	-6.989	3.476	19	-2.011	0.059
	4	0.851	1.752	19	0.486	0.633
	5	0.000	0.071	19	0.000	1.000
	6	-17.019	3.770	19	-4.514	0.000

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
BLOCK	1	-10.711	-18.374	-3.048
	2	0.000	-0.147	0.147
	3	-6.989	-14.264	0.286
	4	0.851	-2.816	4.518
	5	0.000	-0.149	0.149
	6	-17.019	-24.911	-9.128

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
HEIGHT*FREQUENCY	5	19	20.766	0.000

Next let us take our analysis one step further by allowing the random BLOCK effects to be correlated. This would be a natural model to use, if, for example, the randomness of the BLOCK effect is mainly because of certain aspects of the experiment that may affect all the experimenters during any replication of the experiment. Such effects may be fatigue, or random conditions prevailing during the experiment. However, if we assume that all the experimenters are equally affected by the common condition then

it may be reasonable to assume that the variances are all same and so are the covariances between the pairs of distinct blocks. In other words, the covariance matrix has a compound symmetry structure.

The input is:

```
MIXED
CATEGORY HEIGHT FREQUENCY BLOCK
MODEL Y = INTERCEPT + HEIGHT*FREQUENCY
RANDOM BLOCK / STRUCTURE = CS
ESTIMATE
```

The output is:

Fit Statistics

```
Final L-L      : -91.785
-2L-L         : 183.571
AIC           : 189.571
AIC(Corrected) : 190.771
BIC           : 193.105
```

Estimates of Covariance Components

Random Effect	Description	Estimate
BLOCK	Variance	59.452
	Parameter	
	Compound Symmetry	4.291
Error variance	Variance	57.339
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		53.614	5.028	5	10.663	0.000
HEIGHT*FREQUENCY	0*0	-46.245	4.870	19	-9.495	0.000
	0*1	-33.534	4.870	19	-6.885	0.000
	0*2	-23.786	4.870	19	-4.884	0.000
	1*0	-29.289	4.870	19	-6.014	0.000
	1*1	-24.427	4.870	19	-5.015	0.000
	1*2	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		53.614	43.090	64.137
HEIGHT*FREQUENCY	0*0	-46.245	-56.439	-36.051
	0*1	-33.534	-43.728	-23.340
	0*2	-23.786	-33.980	-13.592
	1*0	-29.289	-39.483	-19.095
	1*1	-24.427	-34.621	-14.233
	1*2	0.000	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
BLOCK	1	-4.932	4.655	19	-1.060	0.303
	2	4.932	4.655	19	1.060	0.303
	3	-2.158	4.655	19	-0.464	0.648
	4	7.398	4.655	19	1.589	0.128
	5	4.624	4.655	19	0.993	0.333
	6	-9.864	4.655	19	-2.119	0.047

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
BLOCK	1	-4.932	-14.675	4.811
	2	4.932	-4.811	14.675
	3	-2.158	-11.900	7.585
	4	7.398	-2.345	17.141
	5	4.624	-5.119	14.366
	6	-9.864	-19.606	-0.121

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
HEIGHT*FREQUENCY	5	19	19.570	0.000

In fact, we can take our exploration even further by allowing an arbitrary covariance structure among the experimenters. There is not much reason to do this for this example. Indeed the number of random effect covariance parameters hikes up from a modest 2 for CS structure to a staggering 21. The model has too many covariance parameters to estimate, which leads to less precise estimation. Yet, being a more general model it surely produces smaller residuals. So there is a tradeoff: which one should we lay more emphasis on, more precise estimates or smaller residuals? The Bayesian Information Criterion (BIC) is designed to resolve this dilemma. It cleverly compares both the issues and strikes up a balance, so that better models have smaller BIC values.

Example 2**Structured Covariance Matrix for Random Errors**

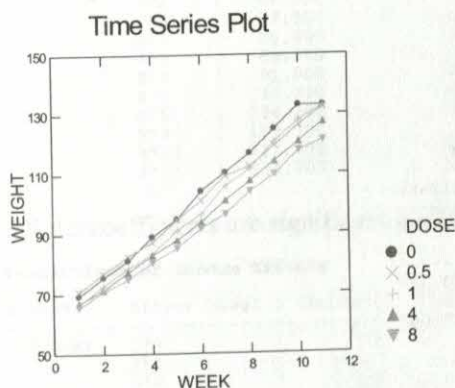
The MIXED command lets the user specify the covariance structure of the random errors. This is particularly useful when we have a mixed model in a time series set up, as in this example. Here we are interested in the effect of the dose of a drug on the growth of rats. We have 5 doses of the drug. Ten rats are assigned to each dose and the weight of each of the 50 rats is observed weekly for 11 weeks. Thus, here the data come from a designed experiment couched in a time series set up. In such a case it may not be a good idea to assume that the random errors are independent. Rather, it is more

natural to assume some stationary time series model for them. But before embarking upon any formal statistical analysis of the data, let us plot the data. We plot the mean body weight against time for the different doses.

The input is:

```
USE RATGROWTH
DOT WEIGHT * WEEK / GROUP = DOSE OVERLAY LINE YMIN=50 YMAX=150
```

The output is:



The plot consists of some nearly parallel straight lines, which leads us to suspect a linear dependence on time, free of any interaction with doses. However, the slight crossovers between the lines near the two extremes doses not allow us to be sure about the absence of the interaction between dose and time. So we use the model

$$Y_{ijk} = \alpha + \beta_{ij} + \varepsilon_{ijk}$$

where $i=1, \dots, 5$, $j=1, \dots, 10$, and $k=1, \dots, 11$. Here y_{ijk} is the weight of the j -th rat under the i -th dose in the k -th week. The rat effect enters through the interaction term β_{ij} . The presence of this interaction, which we treat as a random effect, captures the fact that the rats constitute a random sample from a large population of rats, and that different rats may react to different doses differently. Finally, we model the error as an AR(1) process.

The input is:

```
MIXED
CATEGORY DOSE WEEK RAT
MODEL WEIGHT = INTERCEPT + DOSE + WEEK +
DOSE*WEEK
RANDOM RAT*DOSE
REPEATED / STRUCTURE = AR(1)
ESTIMATE
```

The output is:

Fit Statistics

```
Final L-L      : -2726.998
-2L-L         : 5453.996
AIC           : 5459.996
AIC(Corrected) : 5460.044
BIC           : 5472.609
```

Estimates of Covariance Components

Random Effect	Description	Estimate
RAT*DOSE	Variance	182.892
	Parameter	
Error variance	Variance	2623.384
	Parameter	
	Error	-0.008
	Correlation	
	(AR(1))	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
DOSE*WEEK	0*1	69.393	16.752	450	4.142	0.000
	0*2	75.600	16.752	450	4.513	0.000
	0*3	81.200	16.752	450	4.847	0.000
	0*4	89.100	16.752	450	5.319	0.000
	0*5	95.000	16.752	450	5.671	0.000
	0*6	104.300	16.752	450	6.226	0.000
	0*7	110.500	16.752	450	6.596	0.000
	0*8	116.900	16.752	450	6.978	0.000
	0*9	125.000	16.752	450	7.462	0.000
	0*10	132.900	16.752	450	7.933	0.000
	0*11	141.503	16.752	450	8.447	0.000
	0.5*1	70.910	16.752	450	4.233	0.000
	0.5*2	76.900	16.752	450	4.591	0.000
	0.5*3	82.700	16.752	450	4.937	0.000
	0.5*4	87.300	16.752	450	5.211	0.000
	0.5*5	94.800	16.752	450	5.659	0.000
	0.5*6	101.200	16.752	450	6.041	0.000
	0.5*7	109.300	16.752	450	6.525	0.000
	0.5*8	116.400	16.752	450	6.948	0.000
	0.5*9	122.700	16.752	450	7.325	0.000
	0.5*10	133.400	16.752	450	7.963	0.000
	0.5*11	140.205	16.752	450	8.369	0.000
	1*1	66.492	16.752	450	3.969	0.000
	1*2	72.900	16.752	450	4.352	0.000
	1*3	78.600	16.752	450	4.692	0.000
	1*4	83.500	16.752	450	4.984	0.000
	1*5	91.200	16.752	450	5.444	0.000
	1*6	96.600	16.752	450	5.766	0.000
	1*7	106.000	16.752	450	6.328	0.000

1*8	111.700	16.752	450	6.668	0.000
1*9	120.700	16.752	450	7.205	0.000
1*10	248.300	16.752	450	14.822	0.000
1*11	135.297	16.752	450	8.077	0.000
4*1	67.003	16.752	450	4.000	0.000
4*2	71.400	16.752	450	4.262	0.000
4*3	77.300	16.752	450	4.614	0.000
4*4	83.000	16.752	450	4.955	0.000
4*5	88.000	16.752	450	5.253	0.000
4*6	93.900	16.752	450	5.605	0.000
4*7	101.200	16.752	450	6.041	0.000
4*8	107.700	16.752	450	6.429	0.000
4*9	114.200	16.752	450	6.817	0.000
4*10	120.700	16.752	450	7.205	0.000
4*11	127.199	16.752	450	7.593	0.000
8*1	63.601	16.752	450	3.797	0.000
8*2	68.500	16.752	450	4.089	0.000
8*3	74.900	16.752	450	4.471	0.000
8*4	80.600	16.752	450	4.811	0.000
8*5	85.300	16.752	450	5.092	0.000
8*6	91.900	16.752	450	5.486	0.000
8*7	96.600	16.752	450	5.766	0.000
8*8	104.200	16.752	450	6.220	0.000
8*9	109.700	16.752	450	6.548	0.000
8*10	117.700	16.752	450	7.026	0.000
8*11	124.703	16.752	450	7.444	0.000

All the coefficients are significant, as judged by the small p-values.:

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
<hr/>						
RAT*DOSE	1*0	-7.373	10.530	450	-0.700	0.484
	2*0	1.217	10.531	450	0.116	0.908
	3*0	3.005	10.531	450	0.285	0.776
	4*0	-3.078	10.531	450	-0.292	0.770
	5*0	-9.881	10.531	450	-0.938	0.349
	6*0	-1.411	10.531	450	-0.134	0.893
	7*0	6.191	10.531	450	0.588	0.557
	8*0	2.969	10.531	450	0.282	0.778
	9*0	3.603	10.531	450	0.342	0.732
	10*0	4.758	10.530	450	0.452	0.652
	11*0.5	-4.364	10.530	450	-0.414	0.679
	12*0.5	-0.470	10.531	450	-0.045	0.964
	13*0.5	13.766	10.531	450	1.307	0.192
	14*0.5	7.215	10.531	450	0.685	0.494
	15*0.5	-5.398	10.531	450	-0.513	0.608
	16*0.5	-2.699	10.531	450	-0.256	0.798
	17*0.5	0.363	10.531	450	0.034	0.972
	18*0.5	-1.822	10.531	450	-0.173	0.863
	19*0.5	-0.553	10.531	450	-0.052	0.958
	20*0.5	-6.039	10.530	450	-0.573	0.567
	21*1	-9.127	10.530	450	-0.867	0.387
	22*1	-8.527	10.531	450	-0.810	0.419
	23*1	-3.000	10.531	450	-0.285	0.776
	24*1	-8.564	10.531	450	-0.813	0.417
	25*1	-7.812	10.531	450	-0.742	0.459
	26*1	-5.861	10.531	450	-0.557	0.578
	27*1	3.073	10.531	450	0.292	0.771
	28*1	45.429	10.531	450	4.314	0.000
	29*1	-1.538	10.531	450	-0.146	0.884
	30*1	-4.073	10.530	450	-0.387	0.699
	31*4	-0.142	10.530	450	-0.013	0.989
	32*4	-7.262	10.531	450	-0.690	0.491
	33*4	0.132	10.531	450	0.013	0.990
	34*4	0.492	10.531	450	0.047	0.963
	35*4	1.929	10.531	450	0.183	0.855
	36*4	7.256	10.531	450	0.689	0.491

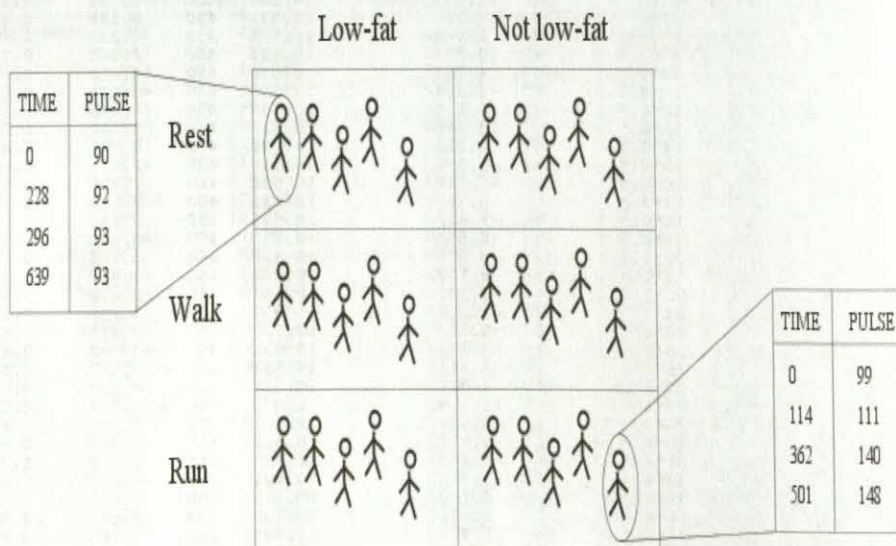
37*4	5.504	10.531	450	0.523	0.601
38*4	-6.421	10.531	450	-0.610	0.542
39*4	-1.183	10.531	450	-0.112	0.911
40*4	-0.305	10.530	450	-0.029	0.977
41*8	-2.095	10.530	450	-0.199	0.842
42*8	1.084	10.531	450	0.103	0.918
43*8	0.489	10.531	450	0.046	0.963
44*8	2.280	10.531	450	0.217	0.829
45*8	-7.066	10.531	450	-0.671	0.503
46*8	8.323	10.531	450	0.790	0.430
47*8	10.354	10.531	450	0.983	0.326
48*8	-11.750	10.531	450	-1.116	0.265
49*8	-3.853	10.531	450	-0.366	0.715
50*8	2.234	10.530	450	0.212	0.832

Almost none of the random coefficients is significant. The only exception is the interaction between rat 28 and dose 1.

Example 3

Repeated Measures Experiment with Covariates

In this example we are interested in analyzing the effect of diet and exercise on pulse rates. We shall consider two types of diet: low-fat and not low-fat; and three types of exercises: at rest, walking leisurely and running. In each of the 6 cells of the resulting two-way layout we assign 5 persons. For each person we measure the pulse rate at 4 different time points. The time points vary from person to person. However, the first measurements are made simultaneously and correspond to time=0.



Let us first plot the pulse rates over time for each of the individuals. The plots suggest that the curves are quadratic in nature. So we shall try to fit a quadratic regression of pulse rate over time:

$$y_{ijkl} = \alpha_{ij} + \beta_{ij}t_{ijkl} + \gamma_{ij}t_{ijkl}^2 + \varepsilon_{ijkl}$$

where y_{ijkl} is the pulse rate at the k -th time point for the l -th person under the (i, j) -th diet-exercise regime. Here $i=1,2, j=1,2,3, k=1,\dots,4$ and $l=1,\dots,5$. Notice that we have allowed the coefficients of the quadratic regression in the 6 cells to depend on the cell. In other words, the coefficients may depend on the particular diet-exercise regime. Now, the diets are qualitatively specified only by their fat contents. So it may be reasonable to assume that the diets used in the experiment are random samples from a large population of diets with the specified fat level. So we further model each of the coefficients α_{ij} , β_{ij} and γ_{ij} as a fixed exercise effect plus a random contribution from the diet. This is an example of multi-level modelling, where the first level consists of modeling the pulse rates in terms of time, and the second level consists of modeling the coefficients in terms of diet and exercise, as shown below.

$$\alpha_{ij} = a + b_j + p_i$$

$$\beta_{ij} = c + d_j + q_i$$

$$\gamma_{ij} = e + f_j + r_i$$

Here p_i , q_i and r_i are the random diet effects in the second level model equations. Combining the two levels, we get the final model

$$y_{ijkl} = (a+b_j) + (c + d_j t_{ijkl}) + (e + f_j t_{ijkl}^2) + (p_i + q_i t_{ijkl} + r_i t_{ijkl}^2) + \varepsilon_{ijkl}$$

Collecting similar terms we see that the right hand side of the model consists of an intercept, the main exercise effect, main time effects (linear and quadratic), and interactions between the times effects and exercise effect.

The input is:

```
USE EXER
MIXED
CATEGORY ID EXERTYPE DIET
MODEL PULSE = INTERCEPT + EXERTYPE + TIME,
        + TIME*EXERTYPE + TIME*TIME,
        + TIME*TIME*EXERTYPE
RANDOM INTERCEPT TIME TIME*TIME/GROUP = DIET
ESTIMATE
```

Notice the GROUP option in the RANDOM line. This fits different random intercepts, linear and quadratic terms for different diets.

The output is:

Fit Statistics

```
Final L-L      : -429.200
-2L-L         : 858.400
AIC           : 866.400
AIC(Corrected) : 866.784
BIC           : 877.165
```

Estimates of Covariance Components

Random Effect	Description	Estimate
Intercept	Variance	12.296
	Parameter	
TIME	Variance	0.000
	Parameter	
TIME*TIME	Variance	0.000
	Parameter	
Error variance	Variance	57.724
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		98.834	3.260	1	30.315	0.021
EXERTYPE	1	-8.715	2.992	104	-2.913	0.004
	2	-5.435	3.007	104	-1.807	0.074
	3	0.000	0.000	.	.	.
TIME		-0.437	284478.623	2	0.000	1.000
TIME*EXERTYPE	TIME*1	0.454	284478.623	104	0.000	1.000
	TIME*2	0.463	284478.623	104	0.000	1.000
	TIME*3	0.570	284478.623	104	0.000	1.000
TIME*TIME		0.001	0.000	.	.	.
TIME*TIME*EXERTYPE	TIME*TIME*1	-0.001	0.000	.	.	.
	TIME*TIME*2	-0.001	0.000	.	.	.
	TIME*TIME*3	-0.001	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		98.834	92.368	105.299
EXERTYPE	1	-8.715	-14.647	-2.783
	2	-5.435	-11.398	0.529
	3	0.000	.	.
TIME		-0.437	-564132.247	564131.372
TIME*EXERTYPE	TIME*1	0.454	-564131.355	564132.264
	TIME*2	0.463	-564131.346	564132.272
	TIME*3	0.570	-564131.240	564132.379
TIME*TIME		0.001	.	.
TIME*TIME*EXERTYPE	TIME*TIME*1	-0.001	.	.
	TIME*TIME*2	-0.001	.	.
	TIME*TIME*3	-0.001	.	.

Predictions of Random Effects

Group	Group Level	Effect	Estimate	Standard Error	df	t	p-value
DIET	1	Intercept	-2.448	2.637	104	-0.928	0.355
DIET	2	Intercept	2.448	2.637	104	0.928	0.355
DIET	1	TIME	-0.004	0.007	104	-0.571	0.570
DIET	2	TIME	0.004	0.007	104	0.571	0.570
DIET	1	TIME*TIME	0.000	0.000	104	-0.336	0.738
DIET	2	TIME*TIME	0.000	0.000	104	0.336	0.738

Confidence Intervals of Random Effects Predictors

Group	Group Level	Effect	Estimate	95.00% Confidence Interval	
				Lower	Upper
DIET	1	Intercept	-2.448	-7.677	2.781
DIET	2	Intercept	2.448	-2.781	7.677
DIET	1	TIME	-0.004	-0.017	0.009
DIET	2	TIME	0.004	-0.009	0.017
DIET	1	TIME*TIME	0.000	0.000	0.000
DIET	2	TIME*TIME	0.000	0.000	0.000

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
EXERTYPE	2	104	0.597	0.552
TIME	1	2	0.000	1.000
TIME*EXERTYPE	3	104	9.182	0.000
TIME*TIME	1	2	0.000	.
TIME*TIME*EXERTYPE	3	104	3.144	0.028

Example 4

Estimation: ML and REML

SYSTAT MIXED allows two different methods to estimate the covariance parameters: Maximum Likelihood (ML) and Residual/Restricted Maximum Likelihood (REML). The more popular choice is REML, which is the default in SYSTAT. For large sample

sizes, the two methods give comparable estimates. The main objection against ML estimators is that they are biased, while REML estimators are unbiased. This data set from Brownlee (1960) pertains to bacteriological testing of milk. Twelve milk samples were tested in all 6 combinations of 2 types of bottles and 3 types of tubes. Ten tests were run on each combination and the response was the number of positive tests in each set of ten.

The input is:

```
USE MILK
MIXED
CATEGORY TUBE$ BOTTLE$
MODEL Y = INTERCEPT + TUBE$ + BOTTLE$ + TUBE$*BOTTLE$
REPEATED / GROUP = SAMPLE
ESTIMATE / METHOD = REML
```

The output is:

Fit Statistics

```
Final L-L      : -131.888
-2L-L         : 263.775
AIC           : 265.775
AIC(Corrected) : 265.838
BIC           : 267.965
```

Estimates of Covariance Components

Random Effect	Description	Estimate
Error variance	Variance	2.542
	Parameter	

Notice that the estimated error variance is 2.542. We shall compare this with the estimate obtained by the ML method later.

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		2.500	0.460	66	5.432	0.000
TUBE\$	A	-1.250	0.651	66	-1.921	0.059
	B	-0.167	0.651	66	-0.256	0.799
	C	0.000	0.000	.	.	.
BOTTLE\$	I	0.167	0.651	66	0.256	0.799
	II	0.000	0.000	.	.	.
TUBE\$*BOTTLE\$	A*I	0.333	0.920	66	0.362	0.718
	A*II	0.000	0.000	.	.	.
	B*I	-0.417	0.920	66	-0.453	0.652
	B*II	0.000	0.000	.	.	.
	C*I	0.000	0.000	.	.	.
	C*II	0.000	0.000	.	.	.

Observe that the effect of Tube A is not significant at 0.05 level (since its p-value is more than 0.05). This inference will change when we use ML later.

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
TUBES	2	66	2.858	0.065
BOTTLES	1	66	0.137	0.713
TUBES*BOTTLES	2	66	0.333	0.718

Next we shall apply ML estimation with the same model.

The input is:

```
USE MILK
MIXED
CATEGORY TUBES$ BOTTLES$
MODEL Y = INTERCEPT + TUBES$ + BOTTLES$ + TUBES*$BOTTLES$
REPEATED / GROUP = SAMPLE
ESTIMATE / METHOD = ML
```

The output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
Error variance	Variance	2.330
	Parameter	

Now the estimated error variance is 2.330, instead of 2.542 as was obtained by the REML method. This is because the REML method adjusts the denominator degrees of freedom to account for the fixed effects. The corrected number of degrees of freedom is less than the raw degrees of freedom used by the ML method. That is why the variance estimate obtained by REML is larger than that obtained by the ML method.

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		2.500	0.441	66	5.674	0.000
TUBES\$	A	-1.250	0.623	66	-2.006	0.049
	B	-0.167	0.623	66	-0.267	0.790
	C	0.000	0.000	.	.	.
BOTTLES\$	I	0.167	0.623	66	0.267	0.790
	II	0.000	0.000	.	.	.
TUBES*\$BOTTLES\$	A*I	0.333	0.881	66	0.378	0.706
	A*II	0.000	0.000	.	.	.
	B*I	-0.417	0.881	66	-0.473	0.638
	B*II	0.000	0.000	.	.	.
	C*I	0.000	0.000	.	.	.
	C*II	0.000	0.000	.	.	.

This table looks similar to what we had using REML estimation earlier. This is because REML is only a little variation of ML. However, the coefficient for Tube A is now significant at 0.05 level (p-value less than 0.05). When we used REML method the same coefficient was not significant at 0.05 level.

Example 5

Hypothesis testing

The data set used here is adapted from Milliken and Johnson (1992). Here we are comparing four different paints. The paints are of two different colors and are manufactured by two different companies. We shall call them Yellow1, Yellow2, White1 and White2, where the 1 and 2 refer to the company. Each paint is applied on three different paving surfaces: Asphalt1, Asphalt2, and Concrete. The response is the life-time measured in weeks. Milliken and Johnson have reported only the cell means and the error sum of squares. The data set has been generated artificially to have the same cell means and error sum of squares as the original data.

We shall fit the following model to this data set:

$$y_{ijk} = \mu_{ij} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

where $i=1,\dots,4$, $j=1,2,3$ and $k=1,2,3$. Here y_{ijk} is the k -th measurement of the life time of the i -th paint as applied on the j -th surface. In this model we shall treat all the effects as fixed. Later we shall see the change introduced by considering some of the effects as random.

The input is:

```
USE PAINTS
MIXED
CATEGORY PAINT$ PAVE$
MODEL Y = INTERCEPT + PAINT$ + PAVE$ + PAINT$*PAVE$
ESTIMATE/METHOD = REML
```

The output is:

Fit Statistics

Final L-L : -75.955
 -2L-L : 151.910
 AIC : 153.910
 AIC(Corrected) : 154.092
 BIC : 155.088

Estimates of Covariance Components

Random Effect	Description	Estimate
Error variance	Variance Parameter	18.961

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		20.000	2.514	24	7.955	0.000
PAINT\$	White1	9.000	3.555	24	2.531	0.018
	White2	16.033	3.555	24	4.510	0.000
	Yellow1	12.000	3.555	24	3.375	0.003
	Yellow2	0.000	0.000	.	.	.
PAVE\$	Asphalt1	7.000	3.555	24	1.969	0.061
	Asphalt2	10.000	3.555	24	2.813	0.010
	Concrete	0.000	0.000	.	.	.
PAINT\$*PAVE\$	White1*Asphalt1	-6.000	5.028	24	-1.193	0.244
	White1*Asphalt2	-11.000	5.028	24	-2.188	0.039
	White1*Concrete	0.000	0.000	.	.	.
	White2*Asphalt1	-9.033	5.028	24	-1.797	0.085
	White2*Asphalt2	-11.033	5.028	24	-2.194	0.038
	White2*Concrete	0.000	0.000	.	.	.
	Yellow1*Asphalt1	-24.000	5.028	24	-4.773	0.000
	Yellow1*Asphalt2	-25.000	5.028	24	-4.972	0.000
	Yellow1*Concrete	0.000	0.000	.	.	.
	Yellow2*Asphalt1	0.000	0.000	.	.	.
	Yellow2*Asphalt2	0.000	0.000	.	.	.
	Yellow2*Concrete	0.000	0.000	.	.	.

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
PAINT\$	3	24	15.790	0.000
PAVE\$	2	24	1.234	0.309
PAINT\$*PAVE\$	6	24	5.638	0.001

Now, a number of hypotheses are of interest. We may be interested in knowing if the Yellow1 paint differs significantly from Yellow2. This corresponds to testing equality of the expectations of the total of the Yellow1 observations and the total of the Yellow2 observations. So we have the hypothesis

$$H_0: 3\alpha_1 - 3\alpha_2 + (\gamma_{11} + \gamma_{12} + \gamma_{13}) - (\gamma_{21} + \gamma_{22} + \gamma_{23}) = 0$$

We can test this hypothesis in SYSTAT as follows:

The input is:

HYPOTHESIS

```
FMATRIX [0,
          3 -3 0 0,
          0 0 0,
          1 1 1,
          -1 -1 -1,
          0 0 0,
          0 0 0]
```

TEST

To understand the F matrix, imagine all the parameters listed in a row in the same order as given in the MODEL line:

$\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \gamma_{11}, \gamma_{12}, \gamma_{13}, \gamma_{21}, \gamma_{22}, \gamma_{23}, \gamma_{31}, \gamma_{32}, \gamma_{33}, \gamma_{41}, \gamma_{42}, \gamma_{43}$

Then, the F matrix is obtained by listing all the null hypothesis coefficients for these parameters.

The output is:

F Matrix

1	2	3	4	5	6
0.000	3.000	-3.000	0.000	0.000	0.000
7	8	9	10	11	12
0.000	0.000	1.000	1.000	1.000	-1.000
13	14	15	16	17	18
-1.000	-1.000	0.000	0.000	0.000	0.000
19	20				
0.000	0.000				

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	24	8.575	0.007

No abbreviation is allowed here. The right-hand side of our H_0 is zero. So you could also write the following:

The input is:

```

HYPOTHESIS
FMATRIX [0,
          1.5 -1.5  0 0,
          0  0  0,
          0.5  0.5  0.5,
          -0.5 -0.5 -0.5,
          0  0  0,
          0  0  0]
TEST

```

because the constant 0.5 just factors out of H_0 .

The output is:

F Matrix

1	2	3	4	5	6
0.000	1.500	-1.500	0.000	0.000	0.000
7	8	9	10	11	12
0.000	0.000	0.500	0.500	0.500	-0.500
13	14	15	16	17	18
-0.500	-0.500	0.000	0.000	0.000	0.000
19	20				
0.000	0.000				

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	24	8.575	0.007

Next let us test if the life-time of the Yellow paints differs from that of the White paints.

The input is:

```

HYPOTHESIS
FMATRIX [0,
          3 3 -3 -3,
          0 0 0,
          1 1 1,
          1 1 1,
          -1 -1 -1,
          -1 -1 -1]
TEST

```

The interaction terms are taken care of by SYSTAT to preserve estimability as mentioned earlier.

The output is:

F Matrix

1	2	3	4	5	6
0.000	3.000	3.000	-3.000	-3.000	0.000
7	8	9	10	11	12
0.000	0.000	1.000	1.000	1.000	1.000
13	14	15	16	17	18
1.000	1.000	-1.000	-1.000	-1.000	-1.000
19	20				
-1.000	-1.000				

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	24	34.339	0.000

Testing between the two types of asphalt can be achieved as follows.

The input is

```

HYPOTHESIS
FMATRIX [0,
          0 0 0 0,
          4 -4, 0,
          1 -1 0,
          1 -1 0,
          1 -1 0,
          1 -1 0]
TEST
  
```

Here the first five 0's keep the μ and α_i 's out of the picture.

The output is:

F Matrix

1	2	3	4	5	6
0.000	0.000	0.000	0.000	0.000	4.000
7	8	9	10	11	12
-4.000	0.000	1.000	-1.000	0.000	1.000

13	14	15	16	17	18
-1.000	0.000	1.000	-1.000	0.000	1.000
19	20				
-1.000	0.000				

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	24	0.316	0.579

So far we have been treating all the effects as fixed. Now let us see what happens if we let the paint effect and the interaction be random. This will be the case, for instance, if the different cans of paints of the same color from the same company show significantly different life-times.

The input is:

```
USE PAINTS
MIXED
CATEGORY PAINT$ PAVE$
MODEL Y = INTERCEPT + PAVE$
RANDOM PAINT$ + PAVE$*PAINT$
ESTIMATE
```

The output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
PAINT\$	Variance	21.389
	Parameter	
PAVE\$*PAINT\$	Variance	29.313
	Parameter	
Error variance	Variance	18.961
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		29.258	3.776	3	7.749	0.004
PAVE\$	Asphalt1	-2.758	4.221	6	-0.653	0.538
	Asphalt2	-1.758	4.221	6	-0.417	0.691
	Concrete	0.000	0.000	.	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
PAINT\$	White1	0.802	3.328	24	0.241	0.812
	White2	4.667	3.328	24	1.402	0.174
	Yellow1	-4.127	3.328	24	-1.240	0.227
	Yellow2	-1.341	3.328	24	-0.403	0.690
PAVE\$*PAINT\$	Asphalt1*White1	2.220	3.886	24	0.571	0.573
	Asphalt1*White2	2.331	3.886	24	0.600	0.554
	Asphalt1*Yellow1	-6.065	3.886	24	-1.561	0.132
	Asphalt1*Yellow2	1.515	3.886	24	0.390	0.700
	Asphalt2*White1	-0.248	3.886	24	-0.064	0.950
	Asphalt2*White2	2.331	3.886	24	0.600	0.554
	Asphalt2*Yellow1	-5.242	3.886	24	-1.349	0.190
	Asphalt2*Yellow2	3.160	3.886	24	0.813	0.424
	Concrete*White1	-0.872	3.886	24	-0.224	0.824
	Concrete*White2	1.734	3.886	24	0.446	0.659
	Concrete*Yellow1	5.651	3.886	24	1.454	0.159
	Concrete*Yellow2	-6.513	3.886	24	-1.676	0.107

Now let us compare the two types of asphalt again.

The input is:

```

HYPOTHESIS
FMATRIX [0,
          1 -1 0]
TEST

```

The output is:

F Matrix

1	2	3	4
0.000	1.000	-1.000	0.000

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	33	0.056	0.814

Observe how the output differs this time. Here we are using the so-called broad inference space, where the random effects are left unspecified. If we want to perform the same comparison but with the effects of the yellow paints held fixed at their currently predicted values, (i.e., if we plan to replicate the experiment with fresh supplies of the white paints, but still use yellow paints of the old standard) then we can do the following:

The input is:

```

HYPOTHESIS
FMATRIX [0,
          1 -1 0]
RMATRIX [3 3 0 0,
          1 1 1,
          1 1 1,
          0 0 0,
          0 0 0]
TEST

```

This is an example of a test performed in intermediate inference space.

The output is:

F Matrix

1	2	3	4
0.000	1.000	-1.000	0.000

R Matrix

1	2	3	4	5	6
3.000	3.000	0.000	0.000	1.000	1.000
7	8	9	10	11	12
1.000	1.000	1.000	1.000	0.000	0.000
13	14	15	16		
0.000	0.000	0.000	0.000		

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	33	0.829	0.369

Broad inference spaces answer most of the needs that occur in real life. One must be very careful to interpret tests performed in other inference spaces.

Example 6

Post hoc tests

The data set used here is adapted from Hand et al. (1994). Data were collected on the genus of flea beetle *Chaetocnema*, which contains three species: *concinna* (Con), *heikertingeri* (Hei), and *heptapotamica* (Hep). Measurements were made on the width and angle of the aedeagus of 74 beetles. The goal of the original study was to form a

classification rule to distinguish the three species. Here we shall analyze if angle has enough information to distinguish among the three classes. First, we fit a one-way ANOVA model.

The input is:

```
USE FLEABEETLE
MIXED
CATEGORY SPECIES$
MODEL ANGLE = INTERCEPT + SPECIES$
ESTIMATE/METHOD = REML
```

The output is:

Fit Statistics

```
Final L-L      : -106.033
-2L-L          : 212.066
AIC            : 214.066
AIC(Corrected) : 214.124
BIC            : 216.329
```

Estimates of Covariance Components

Random Effect	Description	Estimate
Error variance	Variance	1.014
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		10.091	0.215	71	46.996	0.000
SPECIES\$	Con	4.004	0.307	71	13.033	0.000
	Hei	4.199	0.281	71	14.958	0.000
	Hep	0.000	0.000	.	.	.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		10.091	9.663	10.519
SPECIES\$	Con	4.004	3.392	4.617
	Hei	4.199	3.640	4.759
	Hep	0.000	.	.

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
SPECIES\$	2	71	129.633	0.000

Then we test if all the coefficients are the same or not. If this null hypothesis gets accepted, then there is not much hope for us.

The input is:

```
HYPOTHESIS
FMATRIX [0 1 -1 0; 0 1 0 -1]
TEST
```

The output is:

F Matrix

1	2	3	4
0.000	1.000	-1.000	0.000
0.000	1.000	0.000	-1.000

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
2.000	71	129.633	0.000

Notice that the p-value is small and so we can safely reject the null hypothesis. But all that this test tells us is that not all the coefficients are the same. We need to test something stronger: whether all the coefficients are distinct. For this we carry out two pairwise tests using Scheffé's method.

The input is:

```
HYPOTHESIS
PAIRWISE SPECIES$ / SCHEFFE
TEST
```

The output is:

Least squares means for effect SPECIES\$

Level	Estimate	Standard Error	df	t	p-value	95.00% Confidence Interval	
						Lower	Upper
Con	14.095	0.220	71	64.136	0.000	13.657	14.533
Hei	14.290	0.181	71	79.003	0.000	13.930	14.651
Hep	10.091	0.215	71	46.996	0.000	9.663	10.519

Scheffe Test of effect SPECIES\$

SPECIES\$	SPECIES\$	Difference	Standard Error	t	p-value	95.00% Confidence Interval	
						Lower	Upper
Con	Hei	-0.195	0.285	-0.685	0.791	-0.907	0.517
	Hep	4.004	0.307	13.033	0.000	3.236	4.772
Hei	Hep	4.199	0.281	14.958	0.000	3.497	4.901

Example 7

Fine Tuning

In any iterative method there are a number of tuning parameters whose values need to be specified. These include initial values, error tolerance and maximum number of iterations allowed. However, SYSTAT hides most of the details from the user by specifying clever defaults. For instance, before starting the iterations, SYSTAT solves the problem approximately by some simple noniterative method, and then uses the approximate answer as the initial value for the iterative algorithm. However, there may be a rare situation where the user has a better initial value to suggest than what SYSTAT uses by default. In such a case, SYSTAT lets the user override the default. This is an advanced feature, which you will hardly need for most real life data sets. Indeed, we shall use a synthetic data set to illustrate the use of fine tuning.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

where $i=1,\dots,5$, $j=1,2,3$ and $k=1,\dots,100$. We take μ and as fixed, α_i and β_j as random, having independent $N(0,1)$ distribution. The random errors, ε_{ijk} 's are assumed independently distributed as $N(0,0.5)$. We simulate this data set, and estimate the parameters.

The input is:

```
USE SIMUL1
MIXED
CATEGORY I J
MODEL Y = INTERCEPT + I
RANDOM J
ESTIMATE
```

The output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
J	Variance	1.240
	Parameter	
Error variance	Variance	0.502
	Parameter	

Now suppose that some more data are collected. We simulate this fresh data from the same model. The new data set is stored in *SIMUL2*. Rather than analyzing *SIMUL2*

from the scratch, we can specify the estimates from the last analysis as initial values. But first let us see what happens if we do *not* supply the initial values.

The input is:

```
USE SIMUL2
MIXED
CATEGORY I J
MODEL Y = INTERCEPT + I
RANDOM J
ESTIMATE
```

The output is:

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		462.398	
1	ECME	439.514	0.049
2	ECME	438.836	0.002
3	ECME	438.760	0.000
4	ECME	438.746	0.000
5	ECME	438.742	0.000
6	NR	438.741	0.000
7	NR	438.741	0.000

Fit Statistics

```
Final L-L      : -219.371
-2L-L          : 438.741
AIC            : 442.741
AIC(Corrected) : 442.804
BIC            : 449.287
```

Estimates of Covariance Components

Random Effect	Description	Estimate
J	Variance	1.192
	Parameter	
Error variance	Variance	0.491
	Parameter	

Next let us investigate the effect of supplying initial values.

The input is:

```
USE SIMUL2
MIXED
CATEGORY I J
MODEL Y = INTERCEPT + I
RANDOM J
ESTIMATE / GSTART = [1.240 0.502]
```

The output is:

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		438.742	
1	ECME	438.741	0.000
2	ECME	438.741	0.000
3	ECME	438.741	0.000
4	ECME	438.741	0.000
5	NR	438.741	0.000

Fit Statistics

Final L-L	:	-219.371
-2L-L	:	438.741
AIC	:	442.741
AIC(Corrected)	:	442.804
BIC	:	449.287

Estimates of Covariance Components

Random Effect	Description	Estimate
J	Variance	1.192
	Parameter	
Error variance	Variance	0.491
	Parameter	

Notice how the number of iterations have gone down. This is because now the iterations have started already close to the answer. While the advantage in terms of lower computing time requirement may not be important for a high speed computer, this feature may prove useful for online data sets that keep on coming at a high rate. Then this feature may be used to update existing estimators.

References

- Brownlee, K.A. (1960). *Statistical theory and methodology in science and engineering*. New York: John Wiley & Sons.
- CMU:DASL (2005): <http://lib.stat.cmu.edu/DASL/Stories/SteppingandHeartRates.html>
- Hand, D.J., Daly, F., McConway, K., Lunn, D., and Ostrowski, E. (1994). *A handbook of small data sets*. London: Chapman Hall.
- Milliken, G.A., and Johnson, D.E. (1992). *Analysis of messy data, Volume I: Designed experiments*. London: Chapman and Hall.

Hierarchical Linear Mixed Models

Arnab Chakraborty and Ravindra Jore

Hierarchical Linear Mixed Models (HLMM) fits and analyzes mixed models with structured covariance/correlation matrices for random effects and residuals. As in LMM, HLMM also provides Variance Components, Compound Symmetry, Diagonal, and Unstructured as random effects covariance structures. As error covariance structures HLMM provides Variance Components, Compound Symmetry, and Autoregressive(1). You can fit various models like random intercept model, random coefficients model, variance components model, mixed effects ANOVA model, growth-curve model, and models with autocorrelated errors using HLMM. HLMM allows random effects to be both categorical and continuous.

In HLMM, SYSTAT provides two methods to estimate covariance parameters, viz., Maximum Likelihood (ML) and Restricted/Residual Maximum Likelihood (REML). SYSTAT provides the following as default output:

- covariance parameter estimates.
- fixed effect estimates and random effect predictions along with their standard errors, confidence intervals, and t-tests for testing the significance.
- F-ratio tests for fixed effects.
- log-likelihood, Akaike Information Criterion (AIC), Akaike Information Criterion Corrected (AICc), Bayesian Information Criterion (BIC) and iteration history.

HLMM provides save options to save residuals, predictions, model parameter estimates with their standard errors, and other statistics to a new data file you specify.

Statistical Background

A general linear mixed model is a model of the form

$$y = X\beta + Z_1\gamma_1 + \dots + Z_p\gamma_p + \varepsilon$$

where y is the data vector, X and Z_i 's are known matrices (either design matrices or covariate matrices), β is the vector of fixed effects, each γ_i is a vector of random effects, and ε is the random error vector. Here y is a random vector, whose randomness comes partly from the random vector γ and partly from ε . We assume that the random vectors γ_i and ε have independent Gaussian distributions with zero mean and variance matrices having some user-specified structure. Here each γ_i consists of the random coefficients for one random effect. The variance-covariance matrix structure may be different for the different effects. SYSTAT provides the option to specify common covariance parameters for multiple effects. MIXED offers two general estimation techniques: ML and REML. ML method finds the parameter estimates such that -2 log-likelihood is minimum. ML method reports biased estimates since it does not account for the degrees of freedom for the estimation of fixed effects estimates. REML produces unbiased parameter estimates. Both these methods are iterative. The latter produces unbiased estimates. SYSTAT reports the Best Linear Unbiased Estimates (BLUES) of the fixed effects and Best Linear Unbiased Predictors (BLUPs) of the random effects, as well as estimates of the variance parameters. BLUES and BLUPs are accompanied by their estimated standard errors, two-sided 95% confidence intervals, and test of significance.

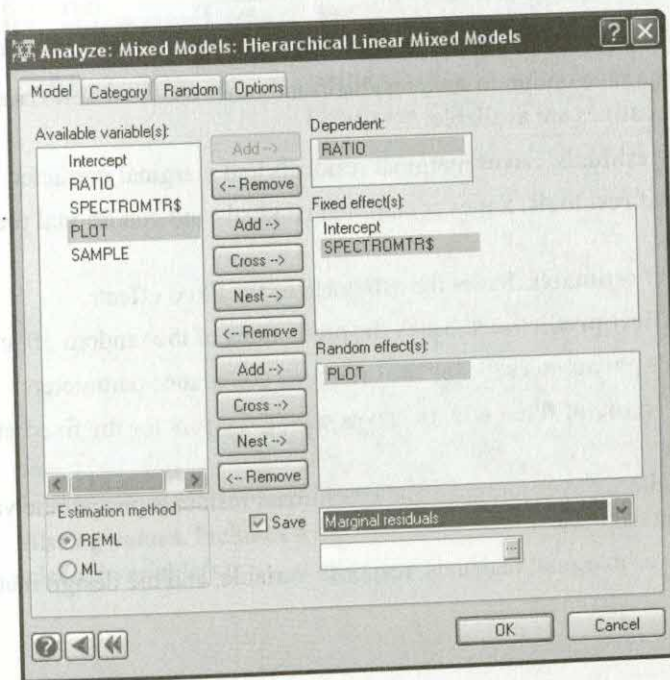
For each model you fit using MIXED, SYSTAT reports log-likelihood, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Akaike Information Criterion Corrected (AICc) for assessing the fit of the model.

Hierarchical Linear Mixed Models in SYSTAT

Model Estimation (in MIXED)

To fit a hierarchical linear mixed model using SYSTAT, from the menus choose:

Analyze
Mixed Models
Hierarchical Linear Mixed Models...



Dependent. Dependent is the variable you want to model. The dependent variable should be continuous and numeric.

Fixed effect(s). Select one or more continuous or categorical variables which you treat as fixed effects. Fixed effects that are not denoted as categorical are considered covariates. If you want crossed or nested effects in your model, you need to build these components using Cross and Nest buttons.

Random effect(s). Select one or more continuous or categorical variables which you treat as random effects. Random effects that are not denoted as categorical are considered covariates. If you want interactions or nested effects in your model, you need to build these components using Cross and Nest buttons. An effect can be fixed as well as random.

Estimation method. Choose one among the available methods to estimate variance components.

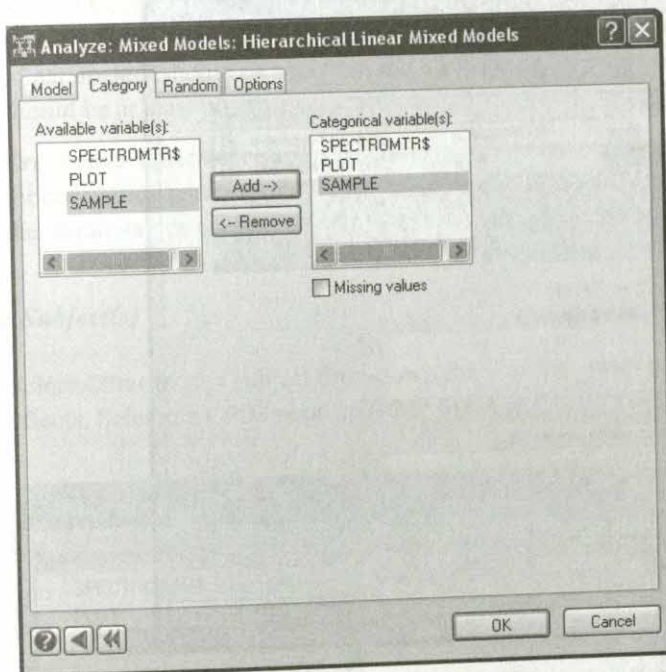
- **REML.** Uses restricted maximum likelihood method to estimate covariance parameters. It is the default method.
- **ML.** Uses maximum likelihood method to estimate covariance parameters.

Save. Check the save option to save residuals and other data to a new data file. The following alternatives are available:

- **Marginal residuals.** Saves marginal residuals and marginal predicted values.
- **Conditional residuals.** Saves conditional residuals and conditional predicted values.
- **Fixed effect estimates.** Saves the estimates of the fixed effects.
- **Random effect predictions.** Saves the predictions of the random effects.
- **Covariance parameters.** Saves the estimated covariance parameters.
- **Standard errors of fixed effects.** Saves standard errors for the fixed effect estimates.
- **Residuals/data.** Saves marginal and conditional residuals plus all the variables in the working data file.
- **Model.** Saves marginal residuals, response variable, and the design matrices.

Category

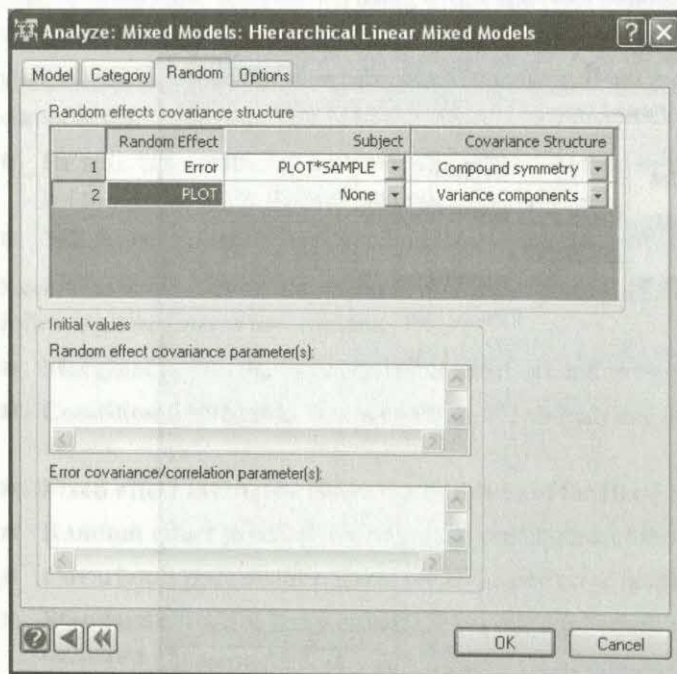
To specify categorical variables, click the Category tab. Select at least one fixed or random effect in Model tab other than intercept to activate this tab.



Missing values. Includes a separate category for cases with a missing value for the selected variable(s).

Random

To specify covariance structures for random effects and errors, click the Random tab.



Random effect. Random effect column lists all the effects denoted as random effects. Error is also listed.

Subject. Specify a subject effect to define hierarchical structure in random effects. Subject defines a block diagonal structure in random effects covariance matrix. That is, the covariance between two subjects is zero. You can define subject effect for errors also.

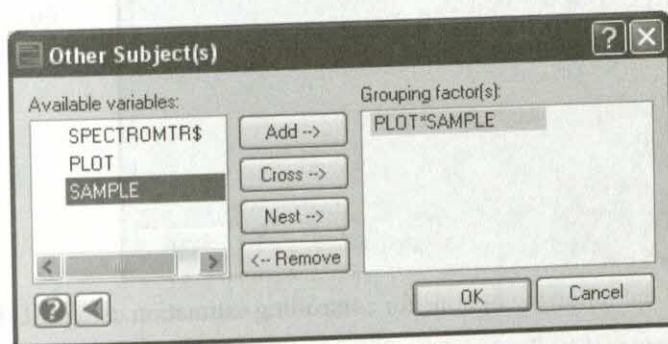
Covariance structure. For a random effect select one of the covariance structures available, viz., Variance components, Diagonal, Compound symmetry, or Unstructured to specify as its covariance structure. For errors, select one of the covariance structures, viz., Variance components, Compound symmetry, or AR(1). The default structure is Variance components.

Random effect covariance parameter(s). Use this option to provide initial values for covariance parameters of random effects. Specify values for each component in the order the effects appear in your model. Separate the values with commas or blanks. You cannot specify initial values for some parameters and leave others blank. Anyhow, SYSTAT computes initial values for all covariance components if you do not specify some/all values. Specify initial values that satisfy parameters constraints. Initial values of parameters should be such that the variance-covariance matrix of random effects should be at least positive semi-definite.

Error covariance/correlation parameter(s). Use this option to provide initial values for correlation parameters. Separate the values with commas or blanks. Make sure that the initial values construct positive-definite error covariance matrix.

Other Subject(s)

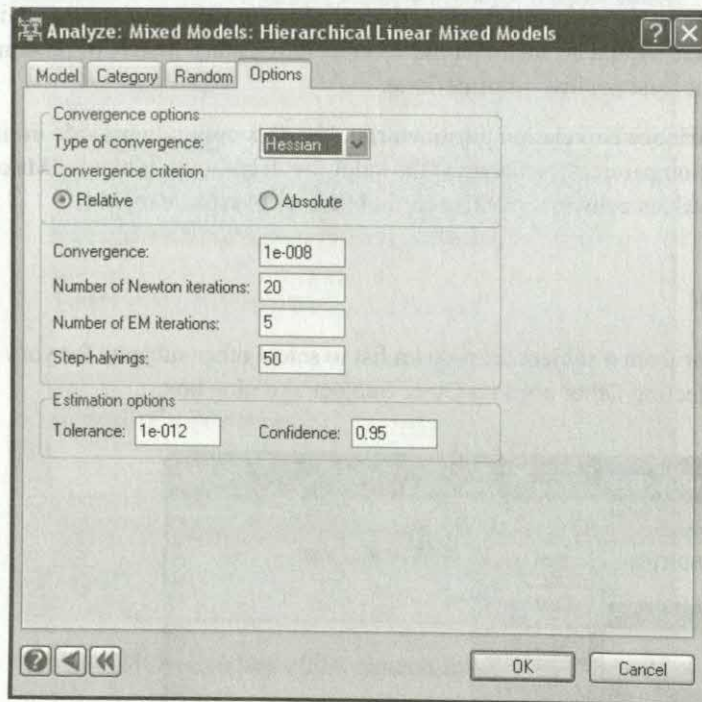
Select Other from a subject drop-down list to select other subjects for your random effects. Selecting Other pops up Other Subject(s) dialog box.



Use Add, Cross, and Nest buttons to build the subjects.

Options

Use Options tab to specify computational controls for ML or REML method of estimation.



SYSTAT offers the following options for controlling estimation using ML/REML:

Type of convergence. Check one of the following options to check convergence.

Three types of convergence checks are available:

- **Hessian.** Uses a quadratic form $g'H^{-1}g$ where g is the gradient vector and H is the hessian matrix.
- **Likelihood.** Uses the difference between log-likelihood at current iteration and the log-likelihood at last iteration.
- **Parameter.** Uses maximum of absolute differences between parameter estimates at current iteration and parameter estimates at last iteration.

Convergence criterion. Two criteria are available:

- **Relative.** Checks relative difference for convergence. That is, convergence checking is done relative to log-likelihood. It is the default option.
- **Absolute.** Tests convergence directly against a value specified.

Convergence. Specify a positive number. MIXED stops iterations when convergence value is less than this number.

Number of Newton iterations. Use this to specify maximum number of Newton-Rapson iterations for fitting your model. The default is 20.

Number of EM iterations. Use this to specify maximum number of EM iterations before going to Newton-Raphson iterations. Sufficient number of EM iterations provide good starting estimates for Newton-Raphson iterations. The default is 5.

Step-halvings. Use this to specify maximum number of step halvings. The default is 50.

Tolerance. A check for near singularity. Use Tolerance to guard against this singularity problem.

Confidence. Specify the confidence coefficient for testing purposes. The default is 0.95.

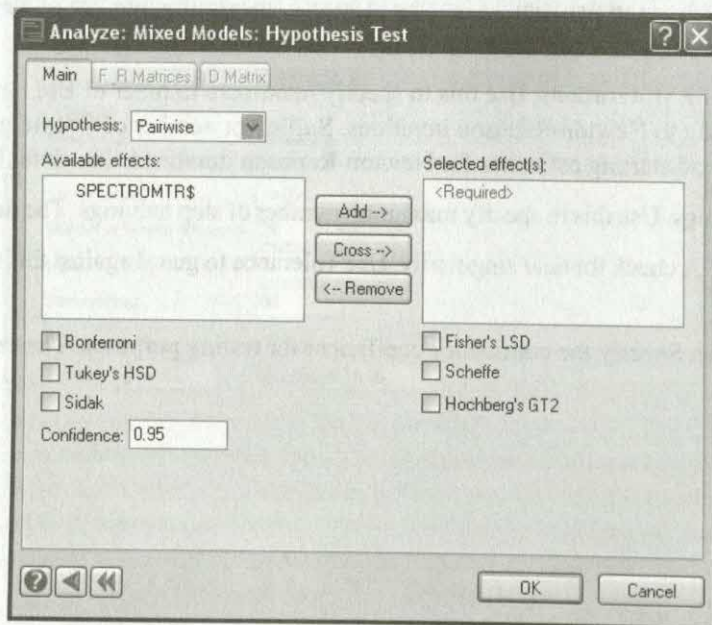
Hypothesis Test

To test hypotheses, from the menu choose:

Analyze

Mixed Models

Hypothesis Test....



You can customize the hypothesis to be tested. You can define contrasts across the categories of a grouping factor:

Hypothesis. Select the type of hypothesis. The following choices are available:

- **F and R Matrices.** Tests the hypotheses corresponding to the F and R Matrices tab.
- **Pairwise.** Compare pairs of groups to determine which pairs differ.

Adjustment method. The following options are available to compute p-value adjustments for multiple comparisons:

- **Bonferroni.** Uses student's t statistics. It sets the family-wise error rate as $(1 - \text{Confidence}) / (\text{Total number of comparisons})$.

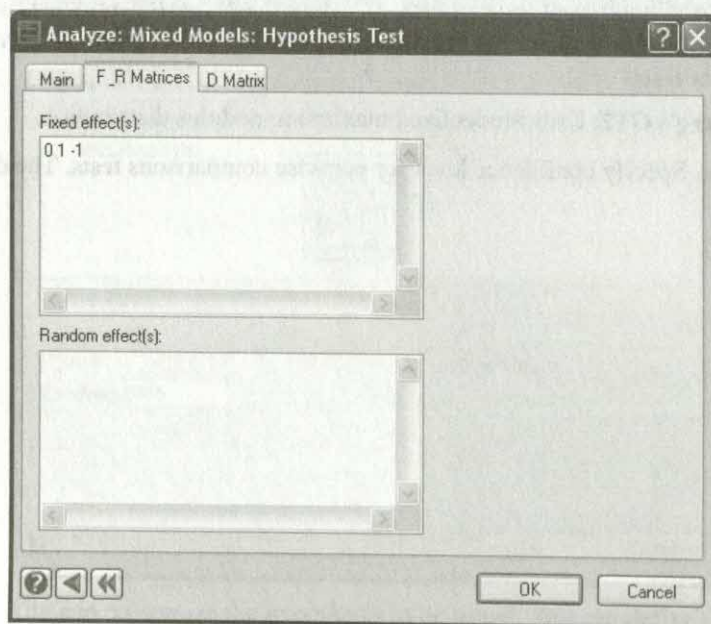
- **Fisher's LSD.** Equivalent to multiple t tests between all pairs of groups. The disadvantage of this test is that no attempt is made to adjust the observed significance level for multiple comparisons.
- **Sidak.** Uses Student's t statistic for pairwise multiple comparisons.
- **Tukey's HSD.** Uses the Studentized range statistic to make all pairwise comparisons. This is the default.
- **Scheffé.** The significance level of Scheffé's test is designed to allow all possible linear combinations of group means to be tested, not just the pairwise comparisons available in this feature. The result is that Scheffé's test is more conservative than the other tests.
- **Hochberg's GT2.** Uses Studentized maximum modulus distribution.

Confidence. Specify confidence level for pairwise comparisons tests. The default is 0.95.

F and R Matrices

To specify **F** and **R** matrices in the Hypothesis drop-down list of the Mixed Models: Hypothesis Test dialog box:

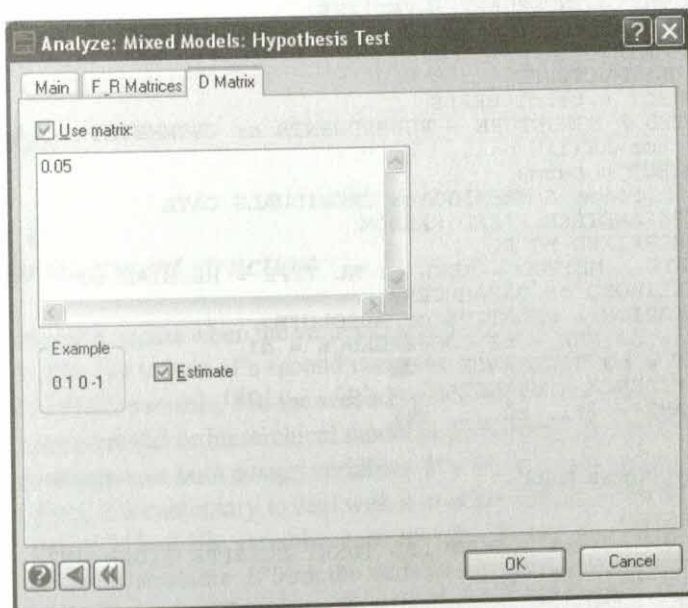
The **F** and **R** Matrices tab gets enabled. *F* and *R* are the matrices of linear weights contrasting the coefficient estimates for fixed and random effects respectively. You can write your hypothesis in terms of the *F* and *R* matrices.



- **Fixed effects.** Specify as many numbers as the dimension of your beta vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.
- **Random effects.** Specify as many numbers as dimension of your gamma vector. In case you specify less, SYSTAT takes the unspecified ones as zero; if you specify more, SYSTAT ignores the extra ones.

D Matrix

D is a null hypothesis vector (by default null vector). The **D** vector, if you use it, must have the same number of rows as the **F** and **R** matrices. To specify a different **D** Matrix, click the D Matrix tab in the Mixed Models: Hypothesis Test dialog box.



Estimate. Check this option for testing significance of contrasts (rows) in **F** and **R** matrices individually. This test reports estimate of the estimable linear parametric function, its standard error and corresponding t-test.

Using Commands

Select the data with USE filename and continue with:

```
MIXED
RESET
CATEGORY grpvarlist / MISS
MODEL var = INTERCEPT + varlist1
RANDOM varlist2 / STRUCTURE = VCOMPONENTS or
  CSYMMETRY
  or UNSTRUCTURED,
  SUBJECT = term1 MEANS
REPEATED / STRUCTURE = VCOMPONENTS or CSYMMETRY
  or AR(1)
  SUBJECT = term2
SAVE filename / MRESIDUALS CRESIDUALS DATA
COVPARAMETERS FIXED RANDOM
ERRORFIXED MODEL
ESTIMATE / METHOD = REML or ML TYPE = HESSIAN or
  LIKELIHOOD or PARAMETERS
CRITERION = RELATIVE or ABSOLUTE
NEM = n1 NNR = n2 CONVERGENCE = d1
HALF = n3 TOLERANCE = d2
CONFIDENCE = d4 GSTART = [g1, ..., gk]
RSTART = [r1, r2, ..., rk]
```

To perform hypothesis tests:

```
HYPOTHESIS
PAIRWISE effect / BONF LSD TUKEY SCHEFFE SIDAK GT2
FMATRIX [matrix]
RMATRIX [matrix]
DMATRIX [matrix]
TEST / CONFI = d1 ESTIMATE
```

Usage Considerations

Types of data. MIXED requires a rectangular data file.

Print options. . MIXED displays covariance parameters and tests of fixed effects for PLENGTH SHORT. For PLENGTH MEDIUM, MIXED adds fixed effects estimates. For PLENGTH LONG, MIXED adds random effects predictions and iteration history.

Quick Graphs. MIXED produces a quick graph of marginal residuals versus marginal predicted values.

Saving files. Several sets of output can be saved to a file. The actual contents of the saved file depend on the analysis. Files may include estimated regression coefficients, model variables, residuals, predicted values, and diagnostic statistics.

BY groups. Each level of any BY variables yields a separate analysis.

Case frequencies. MIXED uses the FREQUENCY variable, if present, to duplicate cases.

Case weights. MIXED uses the values of any WEIGHT variables to weight each case.

Examples

Example 1

Nesting in treatment structure

Nesting occurs when the values of one categorical variable has different interpretations within the values of a second categorical variable. We shall call the first variable the *NESTED* variable, and the second the *NESTING* variable. In statistical jargon the term nested model or hierarchical model is used where the two variables are either both treatments or both design variables. If a treatment effect is nested within a design effect, it is customary to deal with it as either a split-plot design or a repeated measures design. If both the variables correspond to treatments, then we have nesting in treatment structure. If both the variables are design variables, then we have nesting in design structure. This example will illustrate the analysis of the former case using SYSTAT. Nesting in design structure will be dealt with in later examples.

This example is based on a pesticide data set given in Milliken and Johnson (1992). Here we are interested in comparing 11 different brands of pesticides. The first three are produced by company *A*, the next two by company *B*, the next two by company *C*, while company *D* is the manufacturer of the last four brands. To compare these 33 glass containers are used, which are randomly grouped into eleven groups of three. The pesticides are assigned randomly to the groups. The assigned pesticide is applied to the inside of each box in its group. Next a box with 400 mosquitoes and soil with bluegrass is put inside each container. The number of live mosquitoes in each box is counted after 4 hours.

Here we have two treatment effects, company and pesticide. Since the companies produce pesticides of different types, the pesticide effect is nested inside company. For instance, brand 1 of company *A* is different from brand 1 of company *B*. The effects

of the boxes and containers may be absorbed into the random error of the model, since they were assigned at random. A reasonable model to capture this structure is

$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$

where $k=1,2,3$, $j=1,\dots,n_i$, $i=1,\dots,4$, and $n_1=3$, $n_2=2$, $n_3=2$, $n_4=4$. Here y_{ijk} is the observation from the k -th box under the j -th brand from the i -th company.

The input is:

```
USE PESTICIDE
MIXED
  CATEGORY COMPANY$ PESTICIDE
  MODEL Y = INTERCEPT + COMPANY$ + PESTICIDE(COMPANY$)
ESTIMATE
```

Notice that since we are dealing with nesting in treatment structure, we do not have any random effect.

The output is:

```
Dependent Variable : Y
Fixed Factor(s) : COMPANY$, PESTICIDE(COMPANY$)
Fixed Covariate(s) : Intercept
Estimation Method : ANOVA Type III
```

Notice that that SYSTAT is using ANOVA Type III estimation here, even though we have not asked for it. This is because in absence of random effects, this is the default estimation method. As we have seen in earlier examples, the default is REML when random effects are present.

```
Dependent Variable : Y
Fixed Factor(s) : COMPANY$, PESTICIDE(COMPANY$)
Fixed Covariate(s) : Intercept
Estimation Method : ANOVA Type III
```

Dimensions

```
Covariance Parameters : 1
Columns in X : 16
Columns in Z : 0
No. of Observations : 33
```

Error Terms

Effect	Denominator Expression	Error Term
COMPANY\$	MS(Error)	60.545
PESTICIDE(COMPANY\$)	MS(Error)	60.545

Analysis of Variance

Source	Type III SS	Numerator df	Denominator df
COMPANY\$	22515.477	3	22.000
PESTICIDE (COMPANY\$)	1412.583	7	22.000
ERROR	1332.000		22

Analysis of Variance (contd...)

Source	Mean Squares	F-ratio	p-value
COMPANY\$	7505.159	123.959	0.000
PESTICIDE (COMPANY\$)	201.798	3.333	0.014
ERROR	60.545		

Both the p-values are significant (below 0.05, say). We always consider the higher order terms first. Here the highest order term is the nested effect, which is significant. This means that the different pesticides produced by the same company differ significantly among themselves. More specifically, there is at least one company, at least two pesticides of which differ significantly. A latter table in the output will shed more light on this issue. Since the nested term is found significant, we must be careful in our interpretation of the p-value for the main effect due to the companies. Saying something like "The typical pesticide of one company differs from the typical pesticide from other companies" is not entirely correct, since owing to the significant nested effect, there is nothing called a "typical pesticide of a company".

Estimates of Variance Components

Source	Variance Components	Standard Error	Z	p-value	95.00% Confidence Interval	
					Lower	Upper
Error	60.545	15.900	3.808	0.000	29.382	91.709

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		88.667	4.492	22	19.737	0.000
COMPANY\$	A	41.000	6.353	22	6.453	0.000
	B	52.000	6.353	22	8.185	0.000
	C	-4.333	6.353	22	-0.682	0.502
	D	0.000	0.000	.	.	.
PESTICIDE (COMPANY\$)	1 (A)	9.333	6.353	22	1.469	0.156
	2 (A)	-1.333	6.353	22	-0.210	0.836
	3 (A)	0.000	0.000	.	.	.
	1 (B)	1.000	6.353	22	0.157	0.876
	2 (B)	0.000	0.000	.	.	.
	1 (C)	15.000	6.353	22	2.361	0.028
	2 (C)	0.000	0.000	.	.	.
	1 (D)	-13.333	6.353	22	-2.099	0.048
	2 (D)	-19.333	6.353	22	-3.043	0.006
	3 (D)	0.333	6.353	22	0.052	0.959
	4 (D)	0.000	0.000	.	.	.

This table reports the individual t-tests. The rows with dots correspond to the coefficients that are assumed to be 0 as part of the estimability constraint enforced by

SYSTAT. For each company, the last nested coefficient is assumed to be 0. The row 3(A) is one such row. This means that pesticide 3 of company A is considered as the reference for company A. The other pesticides of the same company will be reported with respect to this reference. The insignificant p-value (above 0.05, say) for row 2(A), for example, means that the pesticide 2 of company A does not perform significantly differently from pesticide 3 of the same company.

In fact, the pesticides of company A all perform more or less the same. The pesticides produced by company B are also similar among themselves. The same, however, cannot be said for the other two companies.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		88.667	79.350	97.983
COMPANY\$	A	41.000	27.824	54.176
	B	52.000	38.824	65.176
	C	-4.333	-17.509	8.842
	D	0.000	.	.
PESTICIDE (COMPANY\$)	1 (A)	9.333	-3.842	22.509
	2 (A)	-1.333	-14.509	11.842
	3 (A)	0.000	.	.
	1 (B)	1.000	-12.176	14.176
	2 (B)	0.000	.	.
	1 (C)	15.000	1.824	28.176
	2 (C)	0.000	.	.
	1 (D)	-13.333	-26.509	-0.158
	2 (D)	-19.333	-32.509	-6.158
	3 (D)	0.333	-12.842	13.509
	4 (D)	0.000	.	.

Type III Tests for Fixed Effects

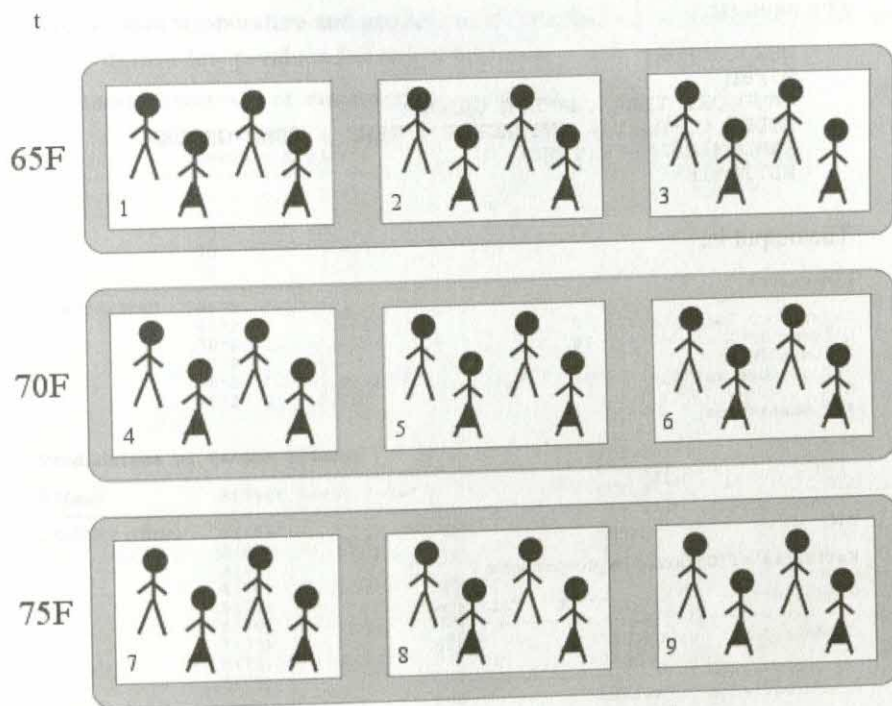
Source	Numerator df	Denominator df	F-ratio	p-value
COMPANY\$	3	22.000	123.959	0.000
PESTICIDE (COMPANY\$)	7	22.000	3.333	0.014

Example 2

Nesting in Design Structure

In this example (Milliken and Johnson, 1992), we consider an experiment to study the effects of temperature on the comfort level of men and women. The experiment was carried out using nine environmental chambers, 18 men and 18 women as follows.

Three different temperatures (65F, 70F and 75F) were assigned to three randomly selected chambers. Two randomly selected men and two randomly selected women were assigned to each chamber. The comfort of each person was measured after three hours in a scale of 1 to 15, where 1=cold, 8=comfortable and 15=hot.



Comfort experiment layout

Here the temperatures are the only treatments, gender and chamber being design effects.

A model for this data set is as follows:

$$y_{ijkl} = \mu + \alpha_i + \gamma_{ik} + \delta_{j(1)} + \varepsilon_{ijkl}$$

where y_{ijkl} is the comfort measurement for the l -th person of k -th gender inside the j -th chamber under temperature i . We shall consider the effects involving chamber as random.

The input is:

```
USE COMFORT
MIXED
CATEGORY TEMP CHAMBER GENDER
MODEL COMFORT = INTERCEPT + TEMP + TEMP*GENDER
RANDOM CHAMBER (TEMP)
ESTIMATE
```

The output is:

Dimensions

```
Covariance Parameters : 2
Columns in X           : 10
Columns in Z           : 9
No. of Observations    : 36
```

Fit Statistics

```
Final L-L      : -61.189
-2L-L         : 122.379
AIC            : 126.379
AIC(Corrected) : 126.823
BIC            : 129.181
```

Estimates of Covariance Components

Random Effect	Description	Estimate
CHAMBER (TEMP)	Variance	2.358
	Parameter	
Error variance	Variance	1.653
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		8.833	1.030	6	8.574	0.000
TEMP	65	-6.000	1.457	6	-4.118	0.006
	70	-0.667	1.457	6	-0.458	0.663
	75	0.000	0.000	.	.	.
TEMP*GENDER	65*1	1.167	0.742	24	1.572	0.129
	65*2	0.000	0.000	.	.	.
	70*1	-2.000	0.742	24	-2.695	0.013
	70*2	0.000	0.000	.	.	.
	75*1	-1.000	0.742	24	-1.347	0.190
	75*2	0.000	0.000	.	.	.

We see that temperature and gender make significant contributions to comfort levels, since quite a few p-values fall below 0.05, say.

Confidence Intervals of Fixed Effects Estimates

Effect	Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
Intercept		8.833	6.707	10.960
TEMP	65	-6.000	-9.007	-2.993
	70	-0.667	-3.674	2.340
	75	0.000	.	.
TEMP*GENDER	65*1	1.167	-0.365	2.699
	65*2	0.000	.	.
	70*1	-2.000	-3.532	-0.468
	70*2	0.000	.	.
	75*1	-1.000	-2.532	0.532
	75*2	0.000	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
CHAMBER (TEMP)	1 (65)	-0.355	1.010	24	-0.351	0.729
	2 (65)	1.135	1.010	24	1.123	0.272
	3 (65)	-0.780	1.010	24	-0.772	0.448
	4 (70)	0.922	1.010	24	0.913	0.371
	5 (70)	-0.780	1.010	24	-0.772	0.448
	6 (70)	-0.142	1.010	24	-0.140	0.890
	7 (75)	2.269	1.010	24	2.246	0.034
	8 (75)	-0.496	1.010	24	-0.491	0.628
	9 (75)	-1.773	1.010	24	-1.755	0.092

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
CHAMBER (TEMP)	1 (65)	-0.355	-2.439	1.730
	2 (65)	1.135	-0.950	3.219
	3 (65)	-0.780	-2.865	1.305
	4 (70)	0.922	-1.163	3.006
	5 (70)	-0.780	-2.865	1.305
	6 (70)	-0.142	-2.227	1.943
	7 (75)	2.269	0.184	4.354
	8 (75)	-0.496	-2.581	1.588
	9 (75)	-1.773	-3.857	0.312

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
TEMP	2	6	7.145	0.026
TEMP*GENDER	3	24	3.849	0.022

None of the design effects are significant (p-values above 0.05, say).

Example 3

Treatment or design?

Sometimes it may be slightly tricky to determine whether nesting occurs in the treatment structure or the design structure. The following example furnishes a case where nesting actually occurs in the treatment structure, though it might appear otherwise at first.

This data set is from Bliss (1967). An experiment was conducted to test the performance of laboratories and technicians to determine the fat content of dried eggs. To this end a single can of dried eggs was stirred well, and 12 samples were drawn. A pair of samples (claimed to be of two "types") was sent to each of six commercial laboratories to be analyzed for fat content. Each laboratory assigned two technicians, who each analyzed both "types".

The dependent variable is the fat content measured by each technician for each sample. The factors in the design are laboratory, technician, and "type". Here "type" is a control effect, while laboratory and technicians are treatment effects (one may treat each technician in each laboratory as a measurement method.).

The experiment has a hierarchical treatment structure. In each of the six laboratories, two technicians examined the fat content of the eggs, but the technicians in each lab were different, so "Technicians" are nested within "Lab". Any technician effect makes sense only within a single laboratory. For instance, looking for a "technician 1" main effect is absurd in this design, because technician 1 of one laboratory may not have anything to do with technician 1 of the other laboratory.

We shall model the data as

$$y_{ijkl} = \mu + \alpha_i + \beta_{j(i)} + \gamma_k + \varepsilon_{ijkl}$$

where y_{ijkl} is the fat content as measured by the j -th technician of the i -th laboratory for the k -th type sample sent to the laboratory. Here we know that the types are actually fakes. So we shall treat the γ_k 's as random effects. All the remaining effects are considered fixed.

The input is:

```
USE EGGS
MIXED
CATEGORY LAB TECHNICIAN$ SAMPLE
MODEL FAT = INTERCEPT + LAB + TECHNICIAN$(LAB)
RANDOM SAMPLE
ESTIMATE
```

The output is

Dimensions

Covariance Parameters : 2
 Columns in X : 19
 Columns in Z : 2
 No. of Observations : 48

Iterations History

Iteration no.	Iteration type	-2L-L	Convergence
0		-50.607	
1	ECME	-50.740	0.003
2	ECME	-50.813	0.001
3	ECME	-50.853	0.001
4	ECME	-50.875	0.000
5	ECME	-50.886	0.000
6	NR	-50.901	0.001
7	NR	-50.901	0.000
8	NR	-50.901	0.000

Fit Statistics

Final L-L : 25.450
 -2L-L : -50.901
 AIC : -46.901
 AIC(Corrected) : -46.537
 BIC : -43.734

Estimates of Covariance Components

Random Effect	Description	Estimate
SAMPLE	Variance Parameter	0.001
Error variance	Variance Parameter	0.009

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		0.175	0.051	1	3.437	0.180
LAB	1	0.548	0.066	35	8.314	0.000
	2	0.140	0.066	35	2.126	0.041
	3	0.270	0.066	35	4.100	0.000
	4	0.203	0.066	35	3.075	0.004
	5	0.173	0.066	35	2.619	0.013
	6	0.000	0.000	.	.	.
TECHNICIAN (LAB)	1 (1)	-0.285	0.066	35	-4.328	0.000
	2 (1)	0.000	0.000	.	.	.
	1 (2)	0.050	0.066	35	0.759	0.453
	2 (2)	0.000	0.000	.	.	.
	1 (3)	-0.075	0.066	35	-1.139	0.263
	2 (3)	0.000	0.000	.	.	.
	1 (4)	-0.002	0.066	35	-0.038	0.970
	2 (4)	0.000	0.000	.	.	.
	1 (5)	0.012	0.066	35	0.190	0.851
	2 (5)	0.000	0.000	.	.	.
	1 (6)	0.185	0.066	35	2.809	0.008
	2 (6)	0.000	0.000	.	.	.

As always, we start by considering the higher order terms first. The p-values less than 0.05 will be considered significant. For laboratories 1 and 6, the two technicians differ significantly. For the other laboratories the technicians have come up with more or less similar measurements. The average measurements of the laboratories are significantly different, as seen by the low p-values for the lab main effect coefficients.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
SAMPLE	1	0.017	0.023	35	0.735	0.467
	2	-0.017	0.023	35	-0.735	0.467

The measurements of different types do not differ significantly. This is not a mere consequence of the fact that the types are actually fakes. A significant difference here would signal some foul play in the entire experiment, e.g., if some dishonest, lazy technicians have provided false measurements and have put significantly different values for the two types just to make the false measurements look "more realistic".

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
LAB	5	35	10.215	0.000
TECHNICIAN (LAB)	6	35	4.755	0.001

This shows something we have already concluded: the measurements from the laboratories differ significantly and so do measurements taken by different technicians within the same laboratory.

Example 4

Nesting versus Crossing

A nested term in a model looks like $\gamma_{j(i)}$ while a crossed (interaction) term looks like γ_{ij} . Both the subscripts are made of the same two indices i and j . Then what is the difference between them? Of course, they have completely different interpretations. However, mathematically, they are essentially the same. This example aims to elucidate this point.

This data set, which is from Beckman et al. (1987), has been analyzed in Hocking (1985) (p 448). This is a study of high efficiency particulate air (HEPA) cartridges. The aim is to compare two types of aerosols used to test the HEPA respirator filters. For this two aerosol types were used with 3 filters from each of two different manufacturers. Since the filters were unique to the manufacturer, we treat the filter effect as nested inside the manufacturer effect.

We shall use the model

$$y_{ijk} = \mu_{ij} + \alpha_{k(i)} + \beta_{jk(i)} + \varepsilon_{ijk}$$

where y_{ijk} is the r -th observation for the k -th filter from the i -th manufacturer on the j -th aerosol. It is assumed that the filters constitute a random sample from a large population of filters. So we treat $\alpha_{k(i)}$'s and $\beta_{jk(i)}$'s as random effects. Beckman et al. (1987) used a simpler model without the interaction term.

The input is:

```
USE AEROSOL
MIXED
CATEGORY MANUFACTURER FILTER AEROSOL
MODEL Y = MANUFACTURER*AEROSOL
RANDOM FILTER (MANUFACTURER) + AEROSOL*FILTER (MANUFACTURER)
ESTIMATE
```

Since this example wants to show the similarity between crossed and nested terms we shall later run the same SYSTAT program with the nested terms replaced by crossed terms. We shall not present the complete output in either case. We shall only present the parts that need to be compared.

The output is:

Estimates of Covariance Components

Random Effect	Description	Estimate
FILTER (MANUFACTURER)	Variance	0.000
	Parameter	
AEROSOL*FILTER (MANUFACTURER)	Variance	0.638
	Parameter	
Error variance	Variance	0.302
	Parameter	

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t
FILTER (MANUFACTURER)	1 (1)	0.000	0.007	24	-0.008
	2 (1)	0.000	0.007	24	0.012
	3 (1)	0.000	0.007	24	-0.004
	1 (2)	0.000	0.007	24	0.004
	2 (2)	0.000	0.007	24	-0.007
	3 (2)	0.000	0.007	24	0.003
AEROSOL*FILTER (MANUFACTURER)	1*1 (1)	0.405	0.520	24	0.779
	1*2 (1)	-0.199	0.520	24	-0.382
	1*3 (1)	-0.207	0.520	24	-0.397
	1*1 (2)	0.553	0.520	24	1.062
	1*2 (2)	-0.256	0.520	24	-0.492
	1*3 (2)	-0.297	0.520	24	-0.571

2*1(1)	-1.161	0.520	24	-2.233
2*2(1)	1.358	0.520	24	2.611
2*3(1)	-0.197	0.520	24	-0.378
2*1(2)	-0.191	0.520	24	-0.367
2*2(2)	-0.361	0.520	24	-0.694
2*3(2)	0.552	0.520	24	1.061

Predictions of Random Effects (contd...)

Effect	Effect Level	p-value
FILTER (MANUFACTURER)	1 (1)	0.994
	2 (1)	0.990
	3 (1)	0.997
	1 (2)	0.997
	2 (2)	0.995
	3 (2)	0.998
AEROSOL*FILTER (MANUFACTURER)	1*1 (1)	0.444
	1*2 (1)	0.706
	1*3 (1)	0.695
	1*1 (2)	0.299
	1*2 (2)	0.627
	1*3 (2)	0.574
	2*1 (1)	0.035
	2*2 (1)	0.015
	2*3 (1)	0.709
	2*1 (2)	0.717
	2*2 (2)	0.495
	2*3 (2)	0.299

Let us make a mental note of some of the values in this table. We shall later compare them with the corresponding values when crossing replaces nesting. The row for 1*1(1) is for aerosol 1 used with filter 1 made by manufacturer 1. The estimate is 0.405. The p-value is 0.442.

Next we replace the nested terms by crossed (interaction) terms.

The input is:

```
USE AEROSOL
MIXED
CATEGORY MANUFACTURER FILTER AEROSOL
MODEL Y = MANUFACTURER*AEROSOL
RANDOM FILTER*MANUFACTURER + AEROSOL*FILTER*MANUFACTURER
ESTIMATE
```

The output is::

Estimates of Covariance Components

Random Effect	Description	Estimate
FILTER*MANUFACTURER	Variance Parameter	0.000
AEROSOL*FILTER*MANUFACTURER	Variance Parameter	0.638
Error variance	Variance Parameter	0.302

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t
AEROSOL*FILTER*MANUFACTURER	1*1*1	0.405	0.520	24	0.779
	1*1*2	0.553	0.520	24	1.062
	1*2*1	-0.199	0.520	24	-0.382
	1*2*2	-0.256	0.520	24	-0.492
	1*3*1	-0.207	0.520	24	-0.397
	1*3*2	-0.297	0.520	24	-0.571
	2*1*1	-1.161	0.520	24	-2.233
	2*1*2	-0.191	0.520	24	-0.367
	2*2*1	1.358	0.520	24	2.611
	2*2*2	-0.361	0.520	24	-0.694
	2*3*1	-0.197	0.520	24	-0.378
	2*3*2	0.552	0.520	24	1.061
FILTER (MANUFACTURER)	1 (1)	0.000	0.007	24	-0.008
	2 (1)	0.000	0.007	24	0.012
	3 (1)	0.000	0.007	24	-0.004
	1 (2)	0.000	0.007	24	0.004
	2 (2)	0.000	0.007	24	-0.007
	3 (2)	0.000	0.007	24	0.003

Predictions of Random Effects (contd...)

Effect	Effect Level	p-value
AEROSOL*FILTER*MANUFACTURER	1*1*1	0.444
	1*1*2	0.299
	1*2*1	0.706
	1*2*2	0.627
	1*3*1	0.695
	1*3*2	0.574
	2*1*1	0.035
	2*1*2	0.717
	2*2*1	0.015
	2*2*2	0.495
	2*3*1	0.709
	2*3*2	0.299
FILTER (MANUFACTURER)	1 (1)	0.994
	2 (1)	0.990
	3 (1)	0.997
	1 (2)	0.997
	2 (2)	0.995
	3 (2)	0.998

We shall now compare this table with the corresponding values from the earlier output. The row for 1*1*1 here corresponds to 1*1(1) earlier. The estimate is again 0.405, and the p-value is 0.442, as before.

Example 5

A Nested-Factorial Model with Case Frequencies

This example focuses on two aspects of the SYSTAT MIXED command: analyzing a nested-factorial model and using case weights. A nested-factorial model is a mixed effects model where both crossing and nesting are present.

Here the data set, which comes from Hocking (1985), is about the concentration of phosphorus in the wash water. The aim of the investigation is to determine how the concentration varies with the types of detergent and washing machines. The experiment was carried out with 4 different types of detergents, 3 different types of machines, and 7 laundromats. The laundromats had different numbers of machines, but each laundromat had only machines of a single type. Thus, laundromats are nested inside machine types. The machines within each laundromat were divided into 4 groups of roughly equal sizes, and the 4 types of detergent were allocated to them. The response is the average amount of phosphorus in grams per liter from daily one-hour samples over a seven day period. The observations have been averaged over all the machines assigned to a single detergent type in each laundromat.

Here the different observations are averages over different numbers of measurements, and contain different amounts of information. We shall take this into account by considering N as the frequencies of cases.

We shall try to fit the following nested-factorial model:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \delta_{k(i)} + \varepsilon_{ijkl}$$

where $r=1, 2$, $k=1, \dots, n_i$, $j=1, \dots, 4$, $i=1, 2, 3$, and $n_1=2$, $n_2=3$, $n_3=2$. Here y_{ijkl} is the r -th observation when the j -th detergent is used in the k -th laundromat with machines of type i .

The input is:

```

USE PHOSPHOR
FREQUENCY N
MIXED
  CATEGORY LAUNDRY DETERG MACHINE
  MODEL Y = INTERCEPT + MACHINE + DETERGENT + MACHINE*DETERG
  RANDOM LAUNDRY(MACHINE) + DETERG*LAUNDRY(MACHINE)
  ESTIMATE

```

This input is somewhat different from others used in this chapter because here we are using case frequencies.

The output is:

Fit Statistics

```

Final L-L      : -25.145
-2L-L         : 50.290
AIC           : 54.290
AIC(Corrected) : 55.213
BIC           : 55.835

```

Estimates of Covariance Components

Random Effect	Description	Estimate
LAUNDRY (MACHINE)	Variance	0.041
	Parameter	
Error variance	Variance	0.471
	Parameter	

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		1.160	0.413	4	2.808	0.048
MACHINE	1	1.840	0.572	4	3.217	0.032
	2	1.304	0.530	4	2.460	0.070
	3	0.000	0.000	.	.	.
DETERG	1	3.090	0.563	12	5.492	0.000
	2	1.846	0.548	12	3.369	0.006
	3	0.590	0.563	12	1.049	0.315
	4	0.000	0.000	.	.	.
MACHINE*DETERG	1*1	-0.312	0.776	12	-0.402	0.695
	1*2	-0.624	0.766	12	-0.815	0.431
	1*3	0.708	0.769	12	0.920	0.375
	1*4	0.000	0.000	.	.	.
	2*1	-2.184	0.719	12	-3.038	0.010
	2*2	-0.643	0.713	12	-0.902	0.385
	2*3	0.999	0.719	12	1.390	0.190
	2*4	0.000	0.000	.	.	.
	3*1	0.000	0.000	.	.	.
	3*2	0.000	0.000	.	.	.
	3*3	0.000	0.000	.	.	.
	3*4	0.000	0.000	.	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
LAUNDRY (MACHINE)	1 (1)	-0.030	0.184	12	-0.164	0.873
	2 (1)	0.030	0.184	12	0.164	0.873
	1 (2)	-0.173	0.178	12	-0.974	0.349
	2 (2)	0.037	0.179	12	0.206	0.840
	3 (2)	0.136	0.178	12	0.766	0.458
	1 (3)	-0.054	0.185	12	-0.293	0.775
	2 (3)	0.054	0.185	12	0.293	0.775

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
MACHINE	2	4	13.670	0.016
DETERG	3	12	19.654	0.000
MACHINE*DETERG	6	12	4.044	0.019

Example 6

Confidence Intervals

The SYSTAT MIXED command can compute confidence intervals for user specified contrasts in various inference spaces (broad, intermediate, and narrow.) In this example we shall explore this with a data set from Brownlee (1960) (Hocking, 1985, p 535). The experiment seeks to compare two different annealing methods for making cans. Three coils of material were selected from the populations of coils made by each of the two methods. A pair of samples was drawn from each of two locations on the coil. The response is the life of the can.

Following Hocking, we shall fit the model

$$y_{ijk} = \mu_{ij} + \alpha_{k(i)} + \beta_{jk(i)} + \varepsilon_{ijk}$$

where i is for method, k for coil within method, and j is for location. Here $\alpha_{k(i)}$ and $\beta_{jk(i)}$ are random effects. Our aim is to produce a 90% confidence interval for the difference between the two methods, i.e., for the contrast.

$$\mu_{11} + \mu_{12} - \mu_{21} - \mu_{22}$$

We shall use the broad inference space (the default.) First we need to fit the model using the SYSTAT input:

```
USE ANNEAL
MIXED
  CATEGORY METHOD LOCATION COIL
  MODEL LIFE = METHOD*LOCATION
  RANDOM COIL(METHOD) LOCATION*COIL(METHOD)
ESTIMATE
```

The output is:

Fit Statistics

```
Final L-L      : -83.771
-2L-L         : 167.542
AIC            : 173.542
AIC(Corrected) : 175.042
BIC           : 176.530
```

Estimates of Covariance Components

Random Effect	Description	Estimate
COIL(METHOD)	Variance	583.194
	Parameter	
LOCATION*COIL(METHOD)	Variance	0.010
	Parameter	
Error variance	Variance	92.539
	Parameter	

Estimates of Fixed Effects

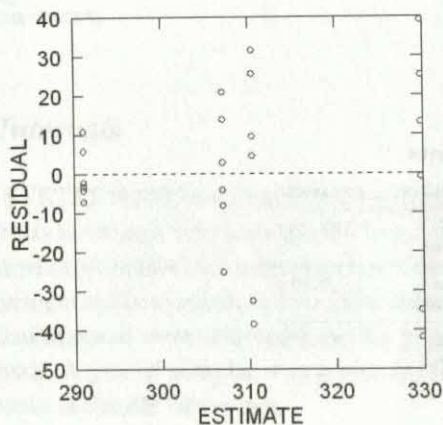
Effect	Level	Estimate	Standard Error	df	t	p-value
METHOD*LOCATION	1*1	329.833	14.485	4	22.770	0.000
	1*2	310.500	14.485	4	21.436	0.000
	2*1	307.167	14.485	4	21.205	0.000
	2*2	291.167	14.485	4	20.101	0.000

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
COIL(METHOD)	1(1)	-35.508	14.465	12	-2.455	0.030
	2(1)	29.176	14.465	12	2.017	0.067
	3(1)	6.332	14.465	12	0.438	0.669
	1(2)	-7.134	14.465	12	-0.493	0.631
	2(2)	6.332	14.465	12	0.438	0.669
	3(2)	0.802	14.465	12	0.055	0.957
LOCATION*COIL(METHOD)	1*1(1)	-0.001	0.099	12	-0.006	0.995
	1*2(1)	0.001	0.099	12	0.006	0.995

1*3(1)	0.000	0.099	12	-0.000	1.000
1*1(2)	-0.001	0.099	12	-0.009	0.993
1*2(2)	0.000	0.099	12	-0.003	0.998
1*3(2)	0.001	0.099	12	0.012	0.991
2*1(1)	0.000	0.099	12	0.000	1.000
2*2(1)	0.000	0.099	12	-0.001	0.999
2*3(1)	0.000	0.099	12	0.001	0.999
2*1(2)	0.001	0.099	12	0.007	0.994
2*2(2)	0.000	0.099	12	0.004	0.997
2*3(2)	-0.001	0.099	12	-0.012	0.991

Plot of residuals against predicted values



Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
METHOD*LOCATION	3	4	7.168	0.044

F Matrix

1	2	3	4
1.000	1.000	-1.000	-1.000

F-ratio Test

Numerator df	Denominator df	F-ratio	p-value
1.000	20	1.091	0.309

Example 7

Nested Random Effects

This example is based on a dataset given by Robinson (1987) (Kuehl, 2000). Two mass spectrometers (SPECTROMTR\$) were compared for accuracy in measuring the ratio of ^{14}N to ^{15}N . Three plots of land (PLOT) treated with ^{15}N were used and from every plot two soil samples (SAMPLE) were taken. Each sample had two observations. The response variable RATIO is the ratio of ^{14}N to ^{15}N multiplied by 1000.

Here PLOT is a random effect and SAMPLE is another random effect nested in PLOT. MACHINES\$ is a fixed effect. That is, the SAMPLE within PLOT is a subject on which repeated observations are taken.

We can perform the mixed models in two ways: Take PLOT as a random effect and SAMPLE as another random effect with grouping factor PLOT, or take PLOT as a random effect and specify compound symmetry structure for errors and take PLOT*SAMPLE interaction as a grouping factor of errors. Both the approaches essentially have the same interpretations. We will use the latter approach.

The input is:

```
USE SPECTROMETERS
MIXED
  CATEGORY PLOT SAMPLE
  MODEL RATIO = INTERCEPT + SPECTROMTR$
  RANDOM PLOT
  REPEATED / STRUCTURE = CS GROUP = PLOT*SAMPLE
ESTIMATE
```

The output is:

Dimensions

```
Covariance Parameters : 2
Columns in X           : 3
Columns in Z           : 3
No. of Observations   : 24
```

Fit Statistics

```
Final L-L             : 48.385
-2L-L                 : -96.770
AIC                   : -90.770
AIC(Corrected)        : -89.437
BIC                   : -87.497
```


Estimates of Covariance Components

Random Effect	Description	Estimate
PLOT	Variance Parameter	0.002
Error variance	Variance Parameter	0.002
	Error Correlation (CS)	0.905

Our subject here is PLOT*SAMPLE. Value 0.905 indicates the higher significance of within-subject correlation.

Estimates of Fixed Effects

Effect	Level	Estimate	Standard Error	df	t	p-value
Intercept		3.863	0.032	2	119.568	0.000
SPECTROMTRS	A	-0.065	0.006	20	-10.543	0.000
	B	0.000	0.000	.	.	.

Predictions of Random Effects

Effect	Effect Level	Estimate	Standard Error	df	t	p-value
PLOT	1	0.030	0.034	20	0.894	0.382
	2	0.009	0.034	20	0.278	0.784
	3	-0.040	0.034	20	-1.172	0.255

Confidence Intervals of Random Effects Predictors

Effect	Effect Level	Estimate	95.00% Confidence Interval	
			Lower	Upper
PLOT	1	0.030	-0.040	0.101
	2	0.009	-0.061	0.080
	3	-0.040	-0.110	0.031

Type III Tests for Fixed Effects

Effect	Numerator df	Denominator df	F-ratio	p-value
SPECTROMTRS	1	20	111.151	0.000

The test of fixed effects indicates that the readings of two spectrometers are significantly different.

References

- Beckman, R. J., Nachtsheim, C. J., and Cook, D. J. (1987). Diagnostics for mixed model analysis of variance. *Technometrics*, 29, 413-426.
- Bliss, C. I. (1967). *Statistics in biology*. McGraw-Hill, New York.
- Brownlee, K.A. (1960). *Statistical theory and methodology in science and engineering*, John Wiley & Sons Inc., New York.
- Hocking, R. R. (1985). *The analysis of linear models*. Wadsworth and Brooks/Cole
- Kuehl, R. O. (2000). *Design of experiments: Statistical principles of research design and analysis*. New York: Duxbury Thomson Learning.
- Milliken, G. A., and Johnson, D. E. (1992). *Analysis of messy data, Volume I: Designed Experiments*. Chapman and Hall.
- Robinson, J. (1967). Incomplete split-plot designs. *Biometrics*, 23, 793-802.

Mixed Regression

Donald Hedeker, Rick Marcantonio, and Michael Pechnyo

Mixed regression estimates models containing combinations of fixed and random effects for response data having a normal distribution. Mixed models, or multilevel models, have also been referred to as "hierarchical linear models" (Bryk and Raudenbush, 2001), "random coefficient models" (deLeeuw and Kreft, 1986), and "variance component models" (Longford, 1993). The implementation here corresponds to the MIXREG program of Hedeker and Gibbons (1996).

These models require a data structure in which observations having a common characteristic can be classified into identifiable groups, known as level-2 units, resulting in nesting of the observations within the level-2 units. Mixed regression uses random effects to account for dependencies in the data due to this nesting structure, allowing simultaneous analysis of individuals and the groups to which the individuals belong. For an individual level-2 unit i , the model for mixed regression is:

$$y_i = \mathbf{W}_i\alpha + \mathbf{X}_i\beta_i + \varepsilon_i$$

where y is the dependent variable, \mathbf{W} is a design matrix for fixed effects, α is a vector of fixed regression parameters, \mathbf{X} is a design matrix for random effects, β is a vector of effects specific to unit i , and ε is a vector of residuals. Models without random effects parallel standard regression models, but use marginal maximum likelihood to derive the parameter estimates instead of least-squares techniques.

Researchers often use mixed regression for the analysis of both clustered and longitudinal data. In clustered data, observations from different subjects are nested within a larger group, such as students within schools; random effects represent differences between the clusters. In contrast, for longitudinal data, observations are nested within each *subject*. In this case, the individual can be viewed as the "cluster", so random effects represent differences between subjects.

Mixed regression, ANOVA, and general linear models can all be used for repeated measures analysis. However, unlike the other two procedures, mixed regression analyzes unbalanced data. Additionally, you can include an autocorrelation structure to model the relationships in the residuals over time.

For each model you fit, the software reports parameter estimates, correlations between estimates, and the intraclass correlation coefficient. You can also view empirical Bayes estimates of the parameters for the random effects. A variety of statistics, including level-1 and level-2 residuals and predicted values can be saved to a file for further analyses and plotting.

Statistical Background

Mixed regression is a modeling technique designed for the analysis of multilevel data. In multilevel data, individual, or level 1, observations can be classified as belonging to known groups, or level 2 units. The data are nested within these groups, leading to a hierarchical structure. The number of observations can vary across level 2 units. The standard data structure appears below:

Level 2 ID	Level 1 ID	Variable			
		Var(1)	Var(2)	...	Var(j)
A	1				
A	2				
:	:				
A	n_A				
B	n_A+1				
B	n_A+2				
:	:				
B	n_A+n_B				
C	n_A+n_B+1				
:	:				

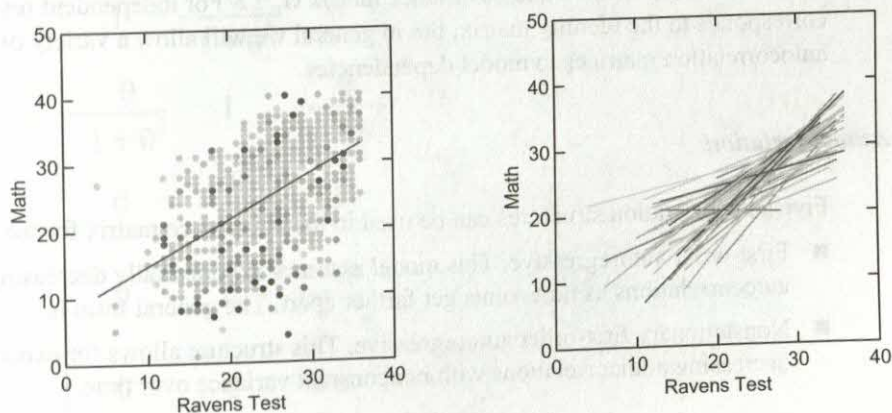
“Clustered” is one common type of multilevel data. In this situation, we have observations from different subjects who can be classified into groups. For example, we may have measurements from students from different classes or schools. Alternatively, we can consider patients nested within doctors, clinics, or hospitals. The goal of the analysis is to examine the effects of variables at both the individual and the group levels.

A second type of multilevel data occurs when collecting repeated measures on each subject. In this case, the level 2 unit corresponds to the person; observations are nested within individuals. The researcher collects measurements over time to examine the effects of time-invariant and time varying variables.

In introducing the basic notions behind mixed regression, we will use a subset of clustered data from the Junior School Project. In the data we examine, roughly 1400 students from 49 schools provided a Ravens test score of ability and a score in mathematics. Because we are using the data for illustrative purposes only, we will not discuss other variables measured, but refer the reader to Mortimore et al. (1988) or the Multilevel Models Project home page (www.ioe.ac.uk/multilevel/) for the complete data file.

Historical Approaches

In the past, several different techniques have been applied to multilevel data. Consider the following two plots:



In the first plot, the regression line ignores any effect due to different schools. Essentially, we are treating all of the data as if it came from one school. Obviously, we are ignoring some potentially important information and are violating the independence assumption inherent in regression. In the second plot, we fit a separate regression line to each school. Interpretation is exceedingly cumbersome and any generalization across schools cannot be made.

Another possibility is to aggregate the level 1 data to the level 2 unit, and perform the analysis at level 2. For the current data, this would involve computing mean scores

for each school, and using the 49 means in the regression. However, the resulting model cannot be applied to individuals and the variation in scores due to individuals is lost. Without that information, relationships appear stronger than they otherwise would.

The General Mixed Regression Model

For level-2 unit i , the general form of the mixed regression model is

$$y_i = W_i\alpha + X_i\beta_i + \varepsilon_i$$

where y represents the response vector, W is a design matrix for fixed effects, α is a vector of fixed regression parameters, X is a design matrix for the random effects, β corresponds to a vector of individual effects, and ε is a residual vector. Terms having a subscript vary across level-2 units.

The random effects have a multivariate normal distribution with mean μ and covariance matrix Σ . The residuals have an independent multivariate normal distribution with mean 0 and covariance matrix $\sigma_\varepsilon^2\Omega$. For independent residuals, Ω corresponds to the identity matrix, but in general we will allow a variety of autocorrelation matrices to model dependencies.

Autocorrelation

Five autocorrelation structures can be used in the covariance matrix for the residuals.

- First-order autoregressive. This model assumes exponentially decreasing autocorrelations as timepoints get farther apart. The general form is:
- Nonstationary first-order autoregressive. This structure allows for exponentially decreasing autocorrelations with nonconstant variance over time.

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- First-order moving average. If θ equals the moving average parameter, the general form of the autocorrelation matrix is:

$$\begin{bmatrix} 1 & -\frac{\theta}{1+\theta^2} & 0 & 0 & 0 \\ -\frac{\theta}{1+\theta^2} & 1 & -\frac{\theta}{1+\theta^2} & 0 & 0 \\ 0 & -\frac{\theta}{1+\theta^2} & 1 & -\frac{\theta}{1+\theta^2} & 0 \\ 0 & 0 & -\frac{\theta}{1+\theta^2} & 1 & -\frac{\theta}{1+\theta^2} \\ 0 & 0 & 0 & -\frac{\theta}{1+\theta^2} & 1 \end{bmatrix}$$

- Autoregressive, moving average (1,1). A combination of a first-order autoregressive process having parameter ϕ with a first order moving average process having parameter θ . The general form equals:

$$\begin{bmatrix} 1 & \omega & \omega\phi & \omega\phi^2 & \omega\phi^3 \\ \omega & 1 & \omega & \omega\phi & \omega\phi^2 \\ \omega\phi & \omega & 1 & \omega & \omega\phi \\ \omega\phi^2 & \omega\phi & \omega & 1 & \omega \\ \omega\phi^3 & \omega\phi^2 & \omega\phi & \omega & 1 \end{bmatrix}$$

where

$$\omega = \frac{(1 - \theta\phi)(\phi - \theta)}{1 - 2\theta\phi + \theta^2}$$

- Toeplitz. General structure having constant autocorrelations along the subdiagonals. The structure equals:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}$$

Fixed Intercept, Fixed Slope Model

The model containing only fixed effects ignores effects due to the nesting of the observations. This model is analogous to the standard linear regression model, but is estimated using marginal maximum likelihood instead of least-squares.

For the JSP data, the estimated parameters and log-likelihood are:

Log Likelihood = -3679.1187

Variable	Estimate	Standardized Error	Z	p-value
INTERCEPT	7.7604	0.7598	10.2132	0.0000
RAVENS_TEST	0.6911	0.0295	23.3884	0.0000
Residual variance:				
	34.4122	1.4326	24.0208	0.0000

Random Intercept, Fixed Slope Model

This model uses the intercept to account for level-2 differences. Each level-2 unit yields a separate intercept. The slope, however, is a constant across all level-2 units. For our data, the following parameter estimates result:

Log Likelihood = -3659.1380

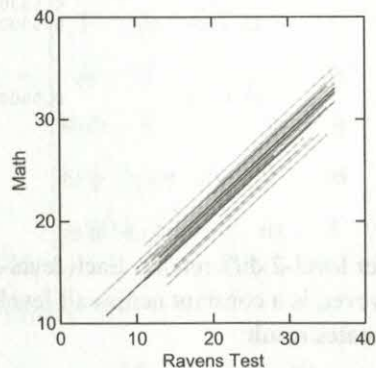
Variable	Estimate	Standardized Error	Z	p-value
INTERCEPT	7.4563	0.7925	9.4091	0.0000
RAVENS_TEST	0.6988	0.0296	23.5898	0.0000
Residual variance:				
	31.9436	1.3581	23.5214	0.0000

Random-effect variance & covariance term(s):

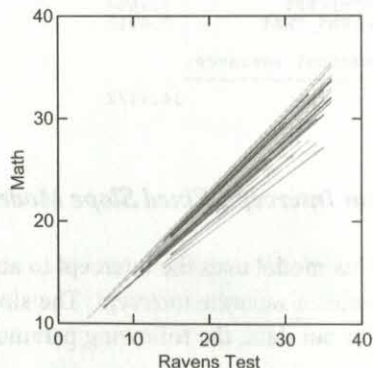
Estimate	
1	INTERCEPT
2.2437	

A plot of this model appears below.

Random Intercept, Fixed Slope



Fixed Intercept, Random Slope



Fixed Intercept, Random Slope Model

This model is used less frequently than the others. In this situation, the slope varies across schools, but the intercept is common. The results for the JSP data are:

Log Likelihood = -3659.3330

Variable	Estimate	Standardized Error	Z	p-value
RAVENS_TEST	0.6909	0.0307	22.4726	0.0000
INTERCEPT	7.6654	0.7567	10.1304	0.0000

Residual variance:

31.9987	1.3602	23.5250	0.0000
---------	--------	---------	--------

Random-effect variance & covariance term(s):

Estimate

		1
	RAVENS_TEST	
1	RAVENS_TEST	0.0032

A plot of this model appears above.

Random Intercept, Random Slope

The completely random model is the most general. In this situation, both the intercept and the slope vary from level-2 unit to level-2 unit. The results for the JSP data are:

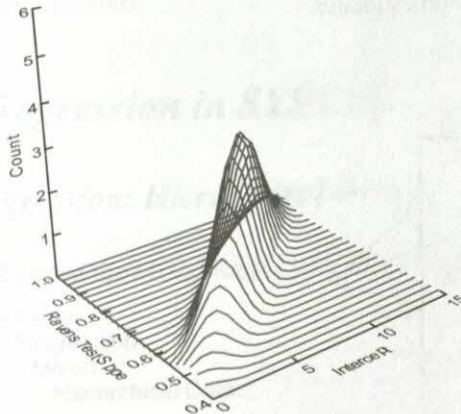
Log Likelihood = -3653.1624

Variable	Estimate	Standardized Error	Z	p-value
INTERCEPT	7.1830	1.0642	6.7494	0.0000
RAVENS_TEST	0.7087	0.0404	17.5327	0.0000
Residual variance:				
	30.8264	1.3361	23.0712	0.0000

Random-effect variance & covariance term(s):

Estimate			
1	2	INTERCEPT	RAVENS_TEST
1	INTERCEPT	24.3553	
2	RAVENS_TEST	-0.8564	0.0332

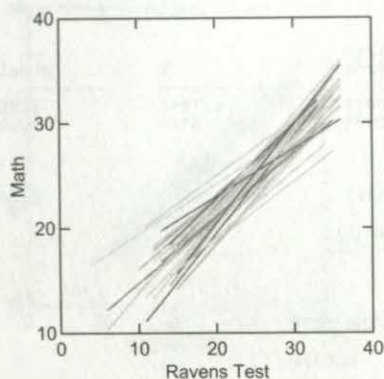
The random effects have a multivariate normal distribution. This distribution appears below.



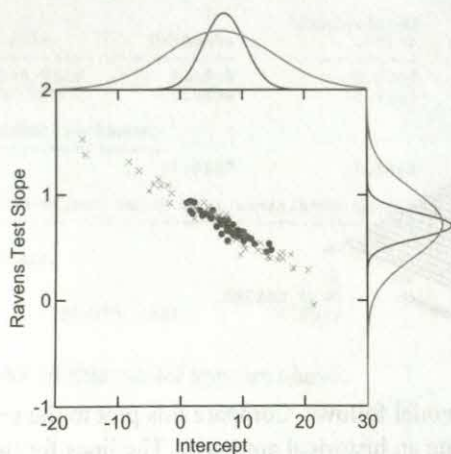
A plot of completely random model follows. Compare this plot to the separate regressions plot used to illustrate an historical approach. The lines for this model are

much more structured.

Random Intercept, Random Slope



The completely random model is similar to computing separate regressions for each level-2 unit. However, mixed regression controls for the group effects in a single model. The following plot compares the least-squares estimates for separate regressions to the mixed regression estimates.



The spread of the intercepts is much less for mixed regression. Similarly, the slopes are less variable. Notice though that both sets of intercepts and both sets of slopes are centered at the same value.

Model Comparisons

To determine whether an effect should be treated as random or fixed, use a likelihood ratio test to compare models that treat the effect each way. The statistic $[-2 \times (\text{the difference between the log-likelihoods})]$ has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated.

For example, comparing the completely fixed model to the random intercept, random slope model yields a statistic of:

$$-2 \times [-3679.1187 - (-3659.1380)] = 39.96$$

The random intercept model adds one parameter to the fixed intercept model, so the degrees of freedom for the test equal 1. The *p-value* for the test is less than 0.0001, indicating that the variability of the random intercept is significant. A single fixed intercept is inadequate for the JSP data. Similar comparisons can be explored for the other models.

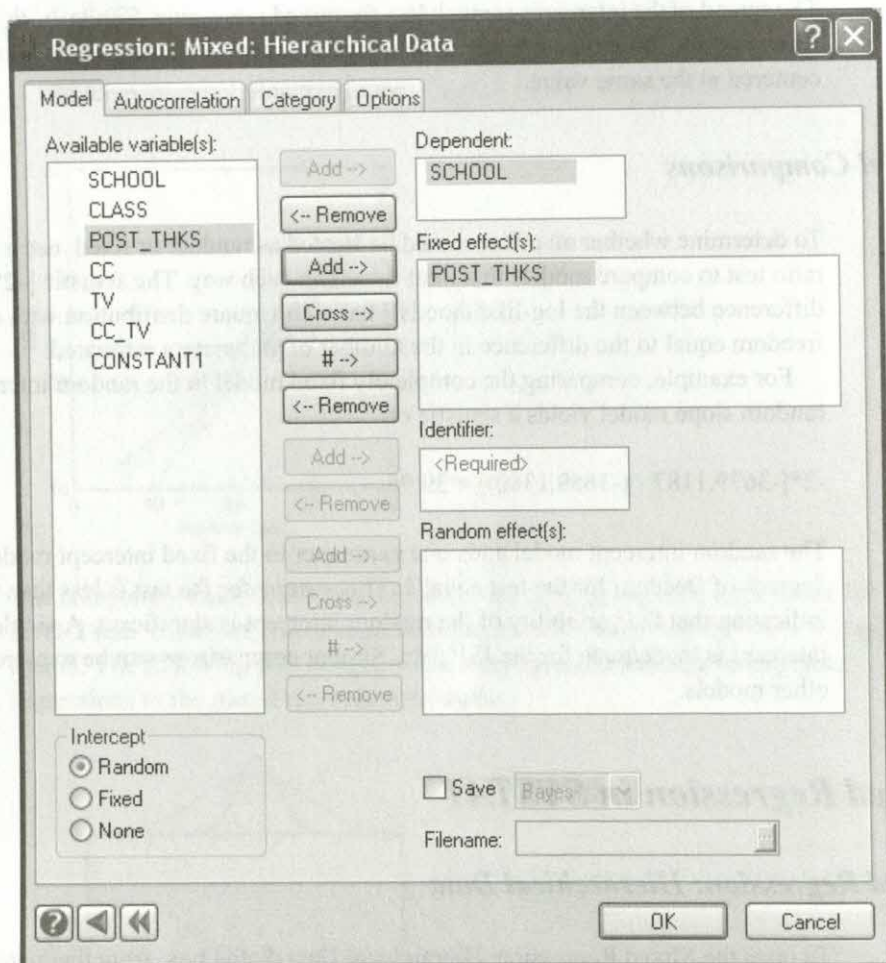
Mixed Regression in SYSTAT

Mixed Regression: Hierarchical Data

To open the Mixed Regression: Hierarchical Data dialog box, from the menus choose:

Analyze
 Regression
 Mixed
 Hierarchical Data...
or

Analyze
 Mixed Models
 Mixed Regression
 Hierarchical Data...



The following options are available:

Dependent. Specify a continuous, numeric variable to be predicted from the fixed and random effects.

Fixed effects. Select one or more continuous or categorical (grouping) variables. Effects corresponding to the selected variables do *not* vary across groups. If you want interactions in your model, you need to build these components using the Cross button.

Identifier. Models containing random effects require an identifier variable to denote the nesting structure. Specify a numeric or string variable that identifies group membership. For cross-sectional data, this variable corresponds to the cluster ID. For longitudinal data, the variable corresponds to the subject ID.

Random effects. Select one or more continuous or categorical (grouping) variables. Effects corresponding to the selected variables vary across groups. If you want interactions in your model, you need to build these components using the Cross button. An effect specified as random is fit as a random effect *and* as a fixed effect. As a result, you cannot fit models in which there are effects that are random but not fixed.

Intercept. Mixed regression models can contain an overall intercept, an intercept that varies across groups, or no intercept.

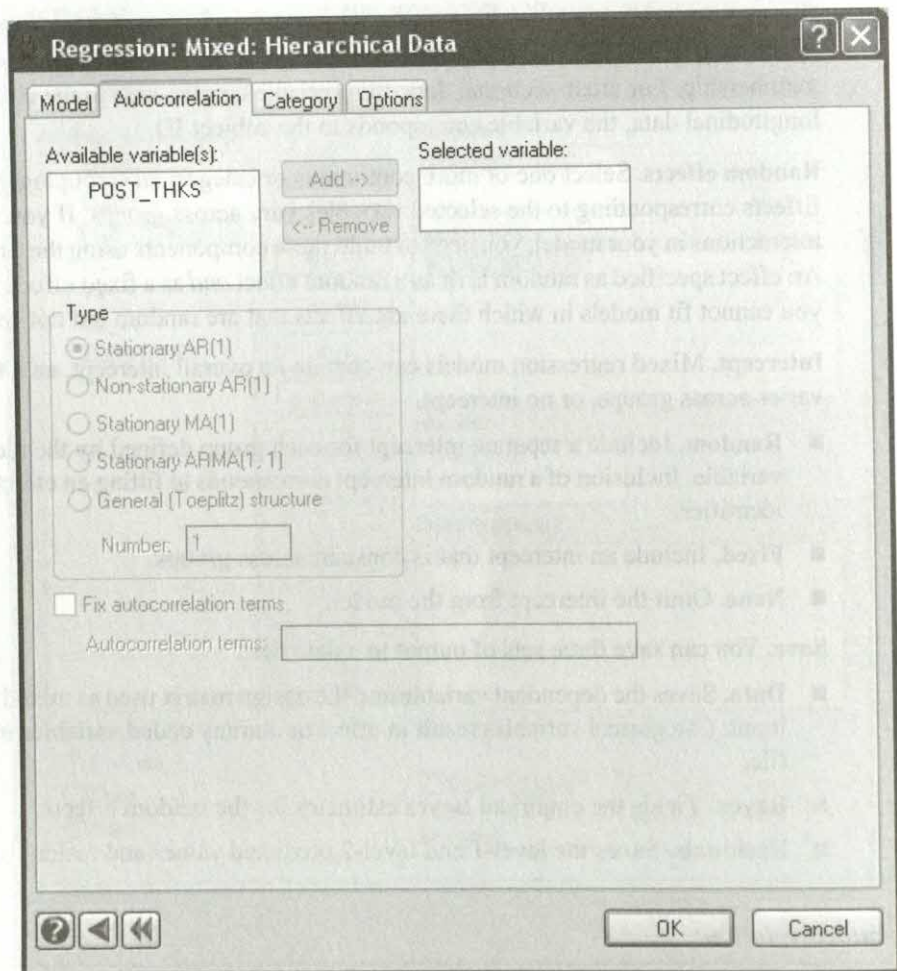
- **Random.** Include a separate intercept for each group defined by the identifier variable. Inclusion of a random intercept corresponds to fitting an effect due to the identifier.
- **Fixed.** Include an intercept that is constant across groups.
- **None.** Omit the intercept from the model.

Save. You can save three sets of output to a data file:

- **Data.** Saves the dependent variable and the design matrix used as mixed regression input. Categorical variables result in effect or dummy coded variables in the saved file.
- **Bayes.** Yields the empirical Bayes estimates for the random effects.
- **Residuals.** Saves the level-1 and level-2 predicted values and residuals.

Autocorrelation

By default, mixed regression assumes the errors are uncorrelated. For longitudinal data, this assumption may be unrealistic, leading to models that include an autocorrelation structure for error to account for dependencies over time. To specify an autocorrelation structure, click Autocorrelation tab in the Mixed : Hierarchical Data dialog box.



Selected variable. Specify a numeric variable that represents measurement occasions, or "time". Typically, this variable is measured in minutes, hours, or days.

The following options are available:

Type. Select one of the following autocorrelation structures for the residual covariance matrix:

- **Stationary AR(1).** Exponentially decreasing correlations as measurement occasions get farther apart.

- **Non-stationary AR(1).** First-order autoregressive process with nonconstant variance over time.
- **Stationary MA(1).** Constant, nonzero correlation between consecutive observations only.
- **Stationary ARMA(1,1).** Structure that is part first-order autoregressive and part first-order moving average.
- **General (Toeplitz) structure.** Constant, nonzero correlation for a specified lag. Enter a lag greater than 0, but less than the maximum number of measurement occasions. For each lag smaller than the specified lag, the correlation is also constant and nonzero, but need not equal the correlation for other lags.

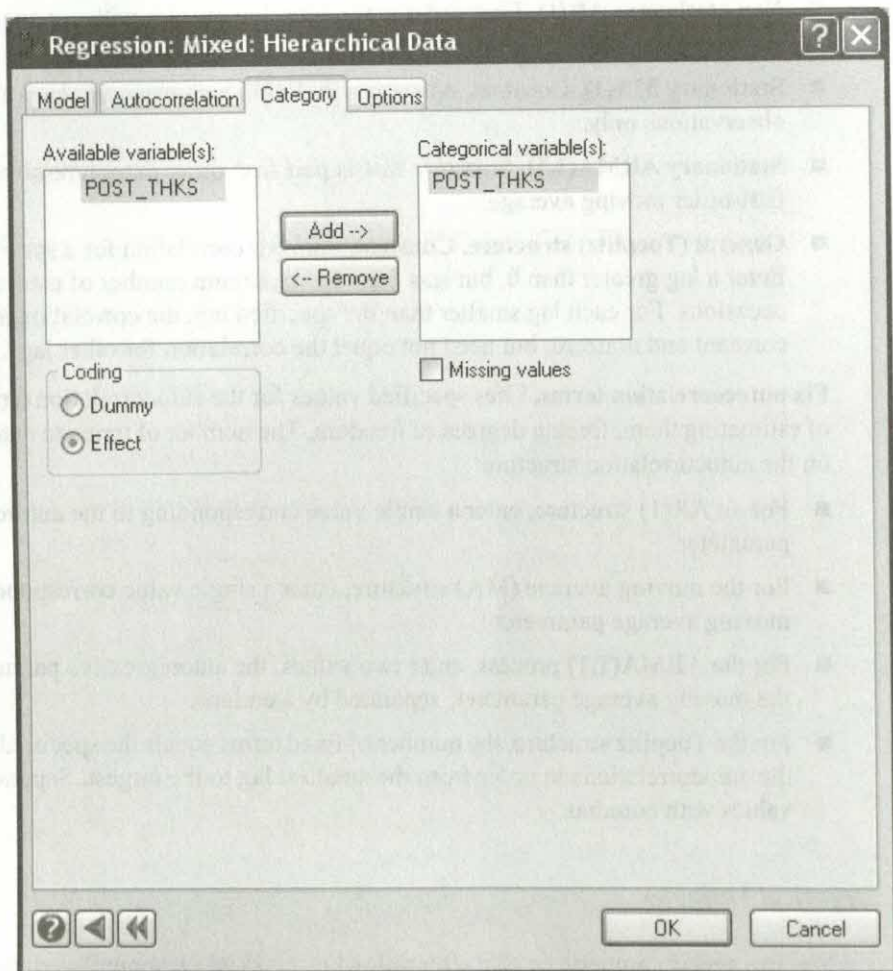
Fix autocorrelation terms. Uses specified values for the autocorrelation terms instead of estimating them, freeing degrees of freedom. The number of terms to enter depends on the autocorrelation structure:

- For an AR(1) structure, enter a single value corresponding to the autoregressive parameter.
- For the moving average (MA) structure, enter a single value corresponding to the moving average parameter
- For the ARMA(1,1) process, enter two values, the autoregressive parameter and the moving average parameter, separated by a comma.
- For the Toeplitz structure, the number of fixed terms equals the specified lag. Enter the autocorrelations in order from the smallest lag to the largest. Separate the values with commas.

Categorical Variables

You can specify numeric or character-valued categorical (grouping) variables that define cells. You want to categorize an independent variable when it has several categories such as education levels, which could be divided into the following categories: less than high school, some high school, finished high school, some college, finished bachelor's degree, finished master's degree, and finished doctorate. On the other hand, a variable such as age in years would not be categorical unless age were broken up into categories such as under 21, 21-65, and over 65.

To define categorical variables, click Category tab in the Mixed : Hierarchical Data dialog box.



The following options are available:

Missing values. Includes a separate category for cases with a missing value for the selected variable(s).

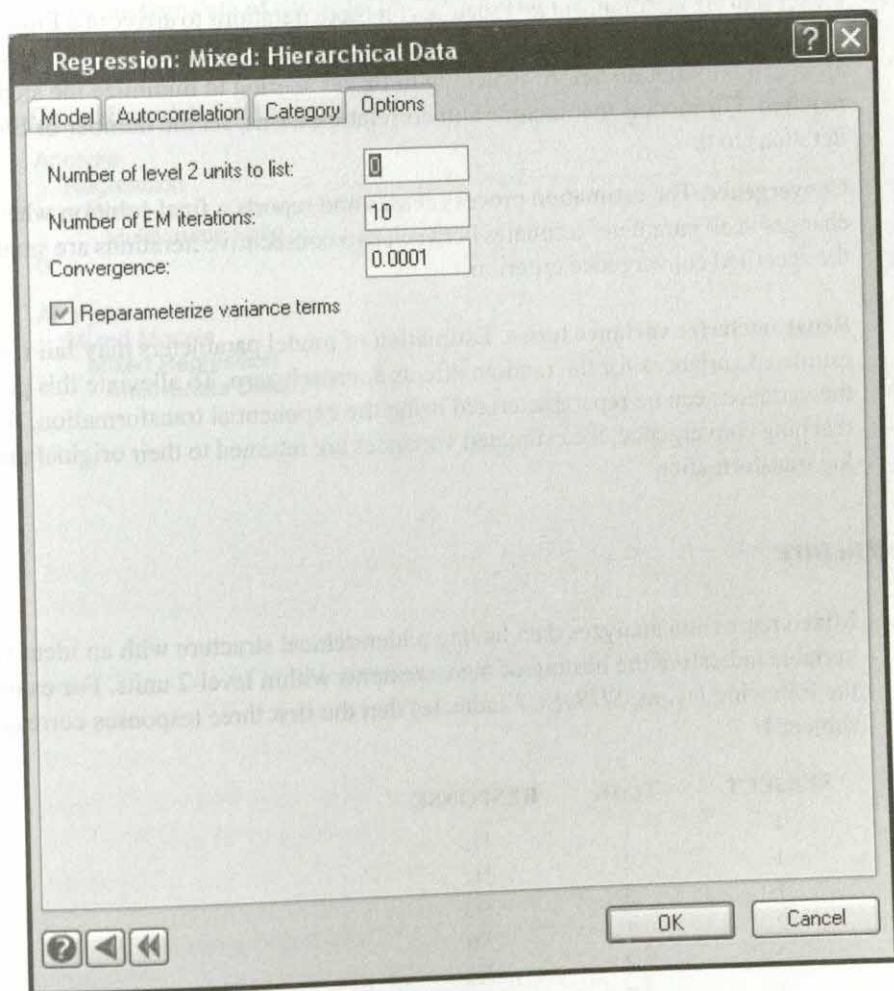
Coding. You can elect to use one of two different coding methods:

- **Dummy.** Produces dummy codes for the design variables instead of effect codes. Coding of dummy variables is the classic analysis of variance parameterization, in which the sum of effects estimated for a classifying variable is 0. If your categorical variable has k categories, $k-1$ dummy variables are created.

- **Effect.** Produces parameter estimates that are differences from group means.

Mixed Regression Options

To specify options for mixed regression models, click Options tab in the Mixed dialog box.



The following options are available:

Number of level 2 units to list. Includes the data for the specified number of groups in the output. When using multivariate data, listing the data along with the output can confirm the conversion to a hierarchical structure.

Number of EM iterations. Specify the number of EM iterations to perform before switching to Fisher scoring. An EM iteration takes much less time to complete than a Fisher scoring iteration, but EM requires far more iterations to arrive at a final solution. In an effort to derive final estimates quickly, estimation uses the EM algorithm to approach the solution before switching to Fisher scoring to minimize the steps required. For models that include autocorrelation terms, set the number of EM iterations to 0.

Convergence. The estimation process ceases and reports a final solution when the changes in all parameter estimates between two consecutive iterations are smaller than the specified convergence criterion.

Reparameterize variance terms. Estimation of model parameters may fail if the estimated variances for the random effects approach zero. To alleviate this problem, the variances can be reparameterized using the exponential transformation. After reaching convergence, the estimated variances are returned to their original units via a log transformation.

Data Structure

Mixed regression analyzes data having a hierarchical structure with an identifier variable indicating the nesting of measurements within level-2 units. For example, in the following layout, *SUBJECT* indicates that the first three responses correspond to subject 1:

SUBJECT	TIME	RESPONSE
1	1	r_{11}
1	2	r_{12}
1	3	r_{13}
2	1	r_{21}
2	2	r_{22}
2	3	r_{23}

An alternative structure, often used for repeated measures data, uses multiple variables to record the responses within each level-2 unit:

SUBJECT	TIME1	TIME2	TIME3
1	r_{11}	r_{12}	r_{13}
2	r_{21}	r_{22}	r_{23}

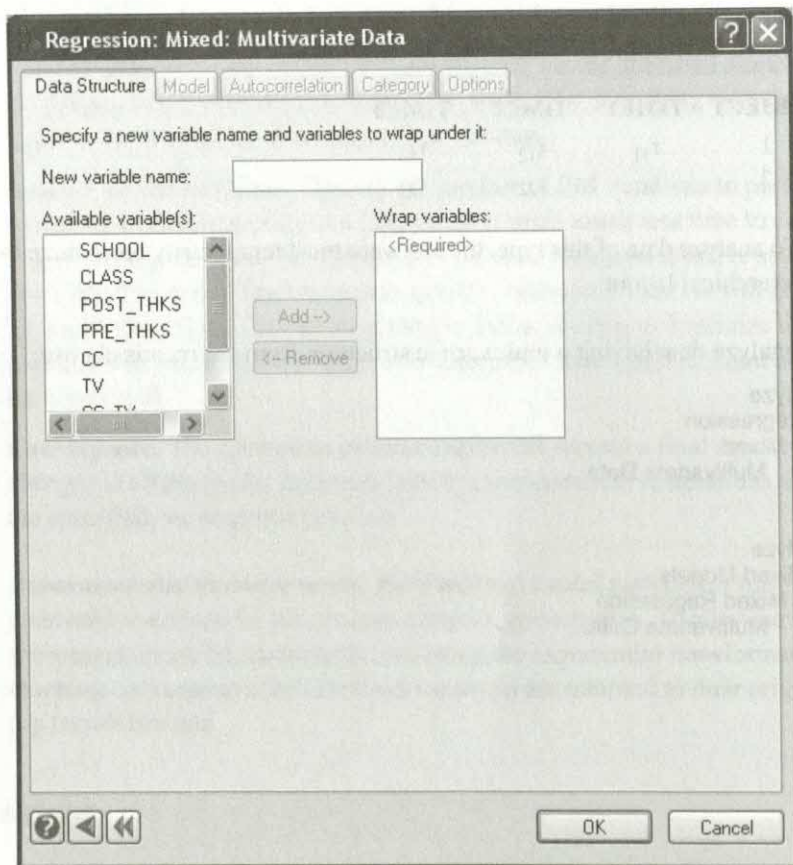
To analyze data of this type, the software must temporarily reorganize the data into a hierarchical layout.

To analyze data having a multivariate structure, from the menus choose:

Analyze
Regression
Mixed
Multivariate Data...

or

Analyze
Mixed Models
Mixed Regression
Multivariate Data...



In the Data Structure dialog box, select the variables to be stacked. Non-selected variables become constants across each associated set of observations in the new data set.

New variable name. Enter a name for the variable under which the selected variables should be stacked. This variable typically corresponds to the dependent variable in the mixed regression model.

The restructured data includes two other new variables, *CASE* and *TRIAL*. *CASE* corresponds to the case number from the multivariate data. *TRIAL* reflects the order of the selected variables.

After restructuring the data, define the mixed regression model. Usually *CASE* denotes the nesting of the observations and should be used as the identifier variable in

your model. For longitudinal data, *TRIAL* represents time and can be used as either a fixed or random effect. In addition, this variable can form the basis of an autocorrelation structure for errors.

Using Commands

First, specify your data with *USE filename*. Continue with:

```
MIX
  RESET
  CONVERT newname = varlist
  MODEL depvar = INTERCEPT + fixedvarlist
  RANDOM INTERCEPT + randomvarlist
  IDENTIFIER var
  AUTO var / TYPE=AR or NAR or MA or ARMA or GEN,
             NUMBER=n FIX=valuelist
  CATEGORY varlist / EFFECT or DUMMY,
                    MISS
  SAVE filename / BAYES RESID DATA
  ESTIMATE / NREC=r NEM=m CONV=crit,
            REPAR=ON or OFF
```

Usage Considerations

Types of data. Mixed regression requires a rectangular data file.

Print options. If *PLENGTH SHORT*, output includes descriptive statistics, parameter estimates, correlations between estimates, and the intraclass correlation coefficient. The *MEDIUM* length adds empirical Bayes estimates to the output. *LONG* adds the iteration history plus variances and covariances for the empirical Bayes estimates. Use *PLENGTH NONE* to suppress all text output.

Quick Graphs. For models containing one random effect, the Quick Graph displays the distribution of the empirical Bayes estimates. Models containing two or more random effects yield a scatterplot matrix of the empirical Bayes estimates.

Saving files. You can save empirical Bayes estimates, residuals with predicted values, or the design matrix. Saved files include effect or dummy coded variables in place of the corresponding categorical variables

BY groups. Mixed regression analyzes data by groups. Your file need not be sorted on the BY variable(s). However, saved files only include results for the first BY group.

Case frequencies. The calculations ignore any *FREQUENCY* variable specifications.

Case weights. Weighting of cases is not available in mixed regression.

Examples

Example 1 Clustered Data in Mixed Regression

To illustrate the use of mixed regression for clustered data, we use data from the Television School and Family Smoking Prevention and Cessation Project. Hedeker and Gibbons (1996) looked at the effects of two factors on tobacco use for students in 28 Los Angeles schools. One factor involved the use of a social-resistance curriculum or not. The other factor was the presence or absence of a television intervention. Crossing these two factors yields four experimental conditions, which were randomly assigned to the schools. Students were measured on tobacco and health knowledge both before and after the introduction of the two factors.

First, we ignore the effects of the nesting within classes by applying a model that includes fixed effects only.

The input is:

```
MIX
USE TVFSP
RESET
MODEL POST_THKS = INTERCEPT+PRE_THKS+CC+CC*TV+TV
SAVE RESIDUALS1 / RESID
PLENGTH SHORT
ESTIMATE

USE RESIDUALS1
LABEL CC / 0='No', 1='Yes'
LABEL TV / 0='No', 1='Yes'
PLOT PRED0*PRE_THKS / OVERLAY GROUP=CC TV SMOOTH=LINEAR,
                        SHORT DASH=9,1,4,10 SIZE= 0,
                        YLAB='Post-intervention THKS',
                        XLAB='Pre-intervention THKS'
```

We save the residuals and predicted values to view the results of the model graphically.

The output is:

Terms in the analysis and names of design matrix columns used for those terms:

CC * TV
FXD4
Perform 10 EM iterations
0 random terms
5 fixed terms

Numbers of Observations

Level 2 observations : 1600
Level 1 observations : 1600

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
POST_THKS	0.000	7.000	2.662	1.383
INTERCEPT	1.000	1.000	1.000	0.000
PRE_THKS	0.000	6.000	2.069	1.260
CC	0.000	1.000	0.477	0.500
TV	0.000	1.000	0.499	0.500
FXD4	0.000	1.000	0.239	0.427

Starting Values

Covariates:
1.661 0.325 0.641 0.199 -0.322
Residual:
1.693

Final Results - MML Estimates

EM Iterations : 10
Fisher Iterations : 2
Total Iterations : 12
Log Likelihood : -2688.962

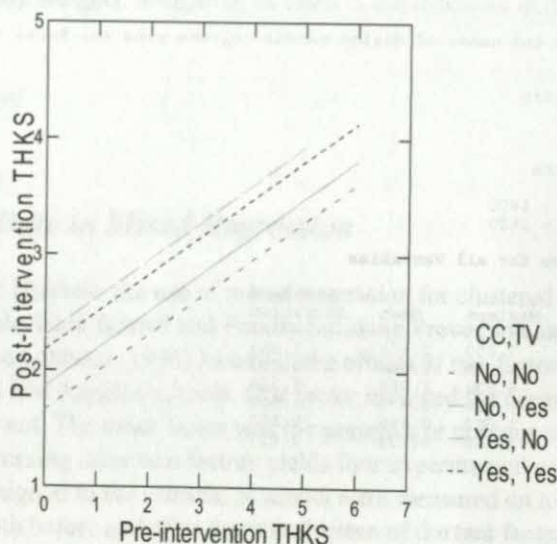
Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	1.661	0.084	19.724	0.000
PRE_THKS	0.325	0.026	12.598	0.000
CC	0.641	0.092	6.966	0.000
TV	0.199	0.090	2.212	0.027
FXD4	-0.322	0.130	-2.473	0.013

Residual Variance

Estimate	Standard Error	Z	p-value
1.688	0.060	28.284	0.000

Correlation of the MML Estimates of the Fixed Terms

	1 INTERCEPT	2 PRE_THKS	3 CC	4 TV	5 FXD4
1 INTERCEPT	1.000				
2 PRE_THKS	-0.659	1.000			
3 CC	-0.536	0.029	1.000		
4 TV	-0.542	0.019	0.486	1.000	
5 FXD4	0.365	0.001	-0.707	-0.690	1.000



The output begins with a note about naming conventions used. The $CC*TV$ interaction receives a root name of FXD because the interaction is a fixed effect. The digit appended to this root corresponds to the position of the effect in the model ($INTERCEPT = 1$; $PRE_THKS = 2$; $CC = 3$; $CC*TV = 4$; $TV = 5$). All subsequent references to $FXD4$ represent the interaction between CC and TV .

Looking at the p-values for the effects, we find a significant effect for all variables in the model. Due to the cross-classification of the CC and TV variables, we actually fit four parallel regression lines.

CC	TV	Regression Line
0	0	$POST_THKS = 1.6613 + 0.3252*PRE_THKS$
0	1	$POST_THKS = 1.8600 + 0.3252*PRE_THKS$
1	0	$POST_THKS = 2.3019 + 0.3252*PRE_THKS$
1	1	$POST_THKS = 2.1790 + 0.3252*PRE_THKS$

These lines correspond to those shown in the plot of the predicted values.

Random Intercept Model

In the TVFSP data, the students can be treated as nested within classes, or as nested within schools. In this example, we consider the effects of the nesting within classes. To account for the data clustering, we use a random intercept.

The input is:

```
MIX
USE TVFSP
RESET
MODEL POST_THKS = PRE_THKS+CC+CC*TV+TV
IDENTIFIER CLASS
RANDOM INTERCEPT
SAVE RESIDUALS1 / RESID
PLENGTH SHORT
ESTIMATE
```

In contrast to the model containing fixed effects only, the random intercept model fits four regression lines *for each school*. Because we treat *PRE_THKS* as a fixed effect, the regression lines are all parallel.

The output is:

```
Terms in the analysis and names of design matrix columns used for those terms:
CC * TV
FXD3
Perform 10 EM iterations
1 random terms
4 fixed terms
```

Numbers of Observations

```
Level 2 observations : 135
Level 1 observations : 1600
```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
POST_THKS	0.000	7.000	2.662	1.383
INTERCEPT	1.000	1.000	1.000	0.000
PRE_THKS	0.000	6.000	2.069	1.260
CC	0.000	1.000	0.477	0.500
TV	0.000	1.000	0.499	0.500
FXD3	0.000	1.000	0.239	0.427

Starting Values

Mean:

1.661

Covariates:

0.325 0.641 0.199 -0.322

Variance Terms:

0.339

Residual:

1.693

Final Results - MML Estimates

EM Iterations : 10
 Fisher Iterations : 7
 Total Iterations : 17
 Log Likelihood : -2679.982

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	1.678	0.099	16.978	0.000
PRE_THKS	0.312	0.026	12.076	0.000
CC	0.633	0.119	5.336	0.000
TV	0.160	0.117	1.368	0.171
FXD3	-0.275	0.168	-1.637	0.102

Residual Variance

Estimate	Standard Error	Z	p-value
1.603	0.059	27.200	0.000

Random-Effect Variance & Covariance Term(s)**Estimate**

	1
	INTERCEPT
1	INTERCEPT
	0.087

Standard Error

	1
	INTERCEPT
1	INTERCEPT
	0.028

Z

	1
	INTERCEPT
1	INTERCEPT
	3.146

p-value

	1
	INTERCEPT
1	INTERCEPT
	0.001

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Calculation of the Intraclass Correlation

Residual Variance : 1.603
 Cluster Variance : 0.087
 Intraclass Correlation : $0.087 / (0.087 + 1.603) = 0.051$

Correlation of the MML Estimates of the Fixed Terms

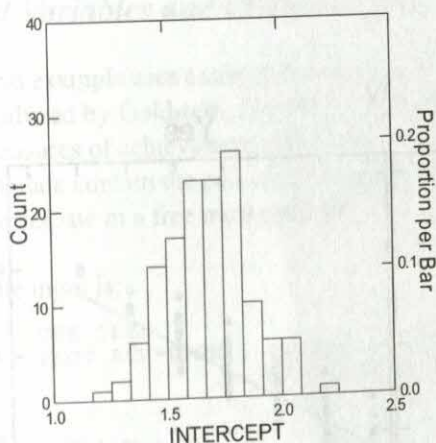
		1	2	3	4	5
		INTERCEPT	PRE_THKS	CC	TV	FXD3
1	INTERCEPT	1.000				
2	PRE_THKS	-0.559	1.000			
3	CC	-0.589	0.028	1.000		

4	TV	-0.593	0.019	0.486	1.000	
5	FXD3	0.408	-0.005	-0.707	-0.695	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2
		VarCov1	Residual
1	VarCov1	1.000	
2	Residual	-0.166	1.000

Empirical Bayes Estimates



The output includes the number of observations at each level for the mixed model. The number of level-2 observations corresponds to the number of groups in the analysis. In this case, the students are nested within 135 classes. The number of level-1 observations indicates the total number of students, 1600. Use PLENGTH MEDIUM to view the number of students within each class.

The individual tests for the parameter estimates indicate significance of the pre-intervention score and of the social-resistance curriculum. In contrast to the fixed effects only model, however, the television intervention and the interaction do not exhibit significant results. Accounting for the classroom effect leads to different conclusions than when we ignore clustering.

The Quick Graph displays the distribution of the empirical Bayes estimates of the intercepts. These values appear to be normally distributed about a value of 1.7. Plotting the predicted values for this model helps to illustrate the effect of fitting a random intercept.

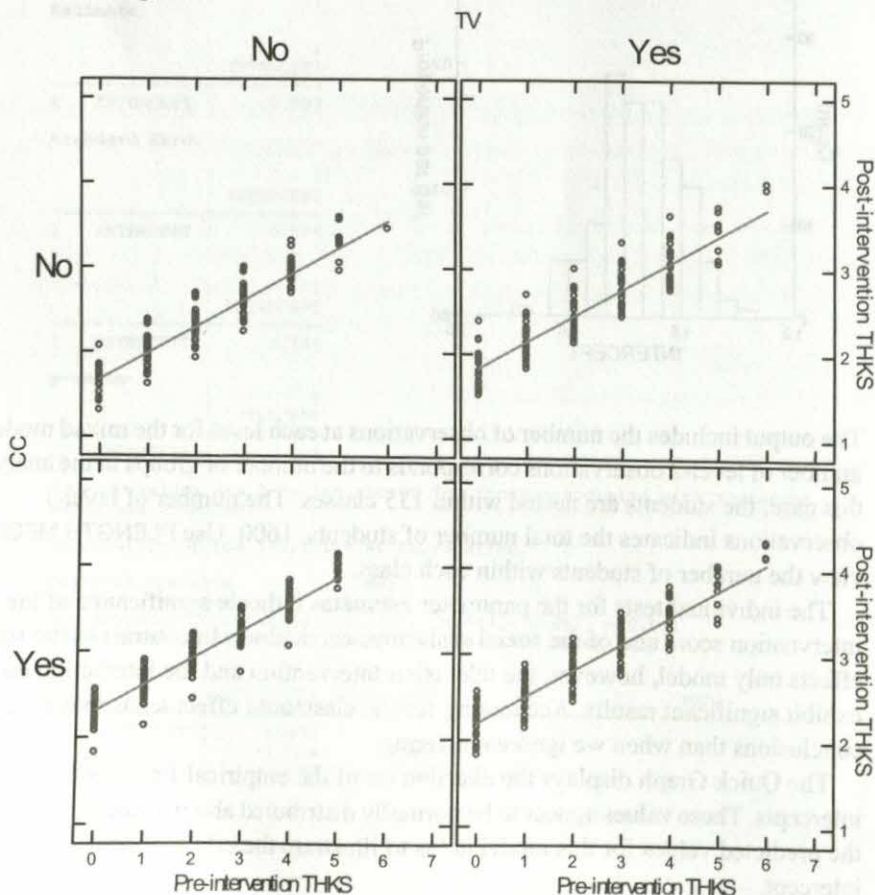
The input is:

```

USE RESIDUALS1
LABEL CC / 0='No', 1='Yes'
LABEL TV / 0='No', 1='Yes'
BEGIN
PLOT PRED1*PRE_THKS / MULTIPLY GROUP=CC TV,
                      YLAB='Post-intervention THKS',
                      XLAB='Pre-intervention THKS',
                      FTITLE=OFF
PLOT PRED0*PRE_THKS / MULTIPLY GROUP=CC TV,
                      SMOOTH=LINEAR SHORT SIZE=0,
                      COLOR=RED YLAB='' XLAB=''
END

```

The output is:



The line indicates the average trend for each *CC*, *TV* combination. The points correspond to the individual trajectory for each class. The fixed effect model generates a single predicted value for each *PRE_THKS* value. The random intercept model generates a predicted value for each *PRE_THKS* value *within each class*. When we allow each class to have a regression line, we eliminate the *TV* and interaction effect present when all classes employed a common regression line for each *CC*TV* combination.

Example 2

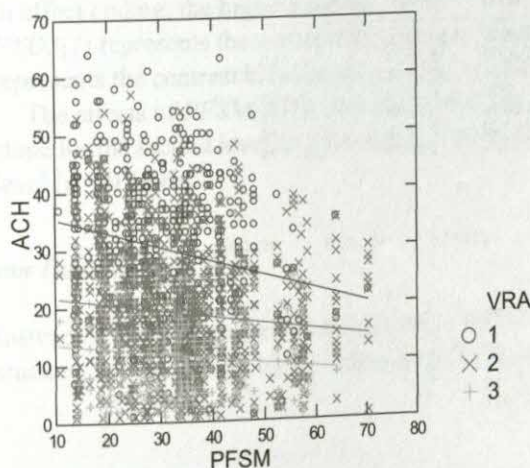
Categorical Variables and Clustered Data

This example uses a subset of data from the Inner London Education Authority (ILEA) analyzed by Goldstein, H.(1987). For 2069 students within 96 schools, we have measures of achievement and a verbal reasoning ability level from 1 to 3. In addition, the data contain the percent of students within each school who are eligible to participate in a free meal program.

The input is:

```
USE ILEA
PLOT ACH*PFSM / GROUP=VRA OVERLAY COLOR=2,1,3,
SMOOTH=LINEAR SHORT
```

The output is:



We begin by fitting a fixed effects only model that includes an intercept and a slope for each level of VRA.

The input is:

```
MIX
USE ILEA
RESET
CATEGORY VRA / EFFECT
MODEL ACH=INTERCEPT+PFISM+VRA+VRA*PFISM
PLENGTH SHORT
ESTIMATE
```

The output is:

Effects coding used for categorical variables in model
Terms in the analysis and names of design matrix columns used for those terms:

```
VRA
FXD3(1)FXD3(2)
VRA * PFISM
FXD4(1)FXD4(2)
Perform 10 EM iterations
0 random terms
6 fixed terms
```

Numbers of Observations

```
Level 2 observations : 2069
Level 1 observations : 2069
```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
ACH	1.000	64.000	20.961	12.282
INTERCEPT	1.000	1.000	1.000	0.000
PFISM	10.760	70.320	31.826	11.636
FXD3(1)	-1.000	1.000	0.109	0.655
FXD3(2)	-1.000	1.000	0.393	0.755
FXD4(1)	-64.000	70.320	2.761	21.770
FXD4(2)	-64.000	70.320	12.605	26.696

Starting Values

```
Covariates:
25.141    -0.153    12.810    -2.219    -0.099    0.035
Residual:
106.861
```

Final Results - MML Estimates

```
EM Iterations      :      10
Fisher Iterations  :       2
Total Iterations   :      12
Log Likelihood     : -7765.476
```

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	25.141	0.767	32.784	0.000
PFSM	-0.153	0.023	-6.744	0.000
FXD3 (1)	12.810	1.082	11.838	0.000
FXD3 (2)	-2.219	0.916	-2.423	0.015
FXD4 (1)	-0.099	0.033	-2.980	0.003
FXD4 (2)	0.035	0.027	1.283	0.200

Residual Variance

Estimate	Standard Error	Z	p-value
106.551	3.313	32.164	0.000

Correlation of the MML Estimates of the Fixed Terms

	1 INTERCEPT	2 PFSM	3 FXD3 (1)	4 FXD3 (2)	5 FXD4 (1)	6 FXD4 (2)
1 INTERCEPT	1.000					
2 PFSM	-0.942	1.000				
3 FXD3 (1)	-0.006	-0.040	1.000			
4 FXD3 (2)	-0.480	0.461	-0.248	1.000		
5 FXD4 (1)	-0.038	0.085	-0.943	0.257	1.000	
6 FXD4 (2)	0.464	-0.498	0.267	-0.940	-0.308	1.000

SYSTAT renames categorical variables according to whether they are fixed (*FXD*) or random (*RND*). To this name, an integer designating the position of the variable in the MODEL command (or the RANDOM command for random effects) is appended. Thus, all instances of *FXD3* refer to *VRA* and instances of *FXD4* refer to the *VRA*PFSM* interaction.

SYSTAT recodes a categorical variable having k levels into $k-1$ dummy variables, using subscripts to denote the contrast represented by the variable. Verbal reasoning ability has three levels, requiring the generation of two dummy variables. Recall that in effect coding, the highest category serves as the reference category. As a result, *FXD3(1)* represents the contrast between the first and third levels of *VRA*. *FXD3(2)* represents the contrast between the second and third *VRA* levels.

The effects of *PFSM*, *VRA*, and the interaction all appear significant. However, the slope for the second level of *VRA* does not appear to differ from the slope for the third level ($p = 0.1995$).

Random Intercepts

Instead of fitting a single line to each *VRA* level, we can account for the clustering of students within schools by including a random intercept.

The input is:

```

MIX
USE ILEA
RESET
CATEGORY VRA / EFFECT
MODEL ACH=PFISM+VRA+VRA*PFISM
RANDOM INTERCEPT
IDENTIFIER SCHOOL
SAVE RESIDUALS1 / RESID
PLENGTH SHORT
ESTIMATE

```

The output is:

Effects coding used for categorical variables in model
 Terms in the analysis and names of design matrix columns used for those terms:

```

VRA
FXD2(1)FXD2(2)
VRA * PFISM
FXD3(1)FXD3(2)
Perform 10 EM iterations
1 random terms
5 fixed terms

```

Numbers of Observations

```

Level 2 observations : 96
Level 1 observations : 2069

```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
ACH	1.000	64.000	20.961	12.282
INTERCEPT	1.000	1.000	1.000	0.000
PFISM	10.760	70.320	31.826	11.636
FXD2(1)	-1.000	1.000	0.109	0.655
FXD2(2)	-1.000	1.000	0.393	0.755
FXD3(1)	-64.000	70.320	2.761	21.770
FXD3(2)	-64.000	70.320	12.605	26.696

Starting Values

```

Mean:
      25.141
Covariates:
      -0.153      12.810      -2.219      -0.099      0.035
Variance Terms:
      21.372
Residual:
      106.861

```

Final Results - MML Estimates

```

EM Iterations      :      10
Fisher Iterations  :       4
Total Iterations   :      14
Log Likelihood     : -7732.113

```

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	25.832	1.179	21.913	0.000
PFSM	-0.171	0.034	-5.036	0.000
FXD2 (1)	12.759	1.072	11.907	0.000
FXD2 (2)	-2.203	0.894	-2.465	0.014
FXD3 (1)	-0.100	0.033	-3.041	0.002
FXD3 (2)	0.033	0.026	1.255	0.209

Residual Variance

Estimate	Standard Error	Z	p-value
98.136	3.122	31.433	0.000

Random-Effect Variance & Covariance Term(s)

Estimate

	1
	INTERCEPT
1	INTERCEPT
	8.947

Standard Error

	1
	INTERCEPT
1	INTERCEPT
	2.021

Z

	1
	INTERCEPT
1	INTERCEPT
	4.428

p-value

	1
	INTERCEPT
1	INTERCEPT
	0.000

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Calculation of the Intraclass Correlation

Residual Variance : 98.136
 Cluster Variance : 8.947
 Intraclass Correlation : $8.947 / (8.947 + 98.136) = 0.084$

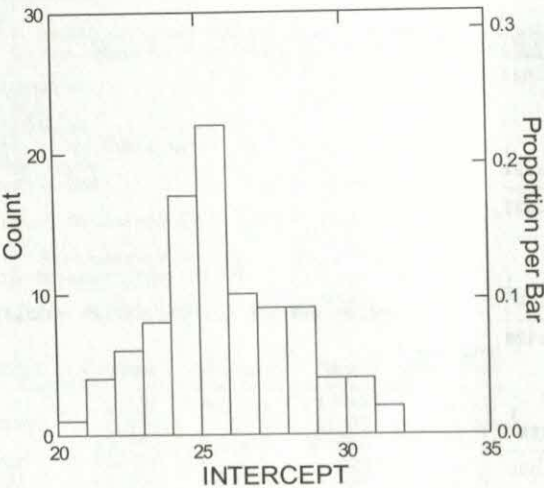
Correlation of the MML Estimates of the Fixed Terms

		1	2	3	4	5	6
		INTERCEPT	PFSM	FXD2 (1)	FXD2 (2)	FXD3 (1)	FXD3 (2)
1	INTERCEPT	1.000					
2	PFSM	-0.940	1.000				
3	FXD2 (1)	0.012	-0.043	1.000			
4	FXD2 (2)	-0.314	0.311	-0.255	1.000		
5	FXD3 (1)	-0.041	0.073	-0.944	0.265	1.000	
6	FXD3 (2)	0.306	-0.337	0.276	-0.941	-0.318	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2
		VarCov1	Residual
1	VarCov1	1.000	
2	Residual	-0.075	1.000

Empirical Bayes Estimates



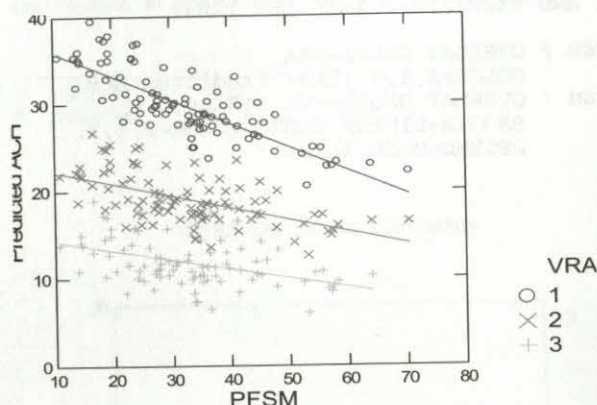
The consequences of fitting the random intercept model can be displayed by plotting the predicted values. For plotting purposes, we need to recreate the original categorical variable, *VRA*, from the corresponding dummy variables.

The input is:

USE RESIDUALS1

```
LET VRA=1
IF FXD2(1) <> 1 AND FXD2(2)=1 THEN LET VRA=2
IF FXD2(1)=-1 AND FXD2(2)=-1 THEN LET VRA=3
BEGIN
PLOT PRED1*PFSM / OVERLAY GROUP=VRA,
                  COLOR=2,1,3 YLAB='Predicted ACH'
PLOT PRED0*PFSM / OVERLAY GROUP=VRA SIZE=0,
                  SMOOTH=LINEAR SHORT COLOR=2,1,3,
                  LEGEND=NONE YLAB=''
END
```

The output is:



In the fixed effect model, every predicted value would lie on the lines in the plot. The random intercept allows each school its own regression line, and thus the predicted values vary by school.

Predicted Values and Confidence Bands

The previous plot showed the scatter of predicted values around the average regression line. The regression line for each school generated the predicted values. Can we see these lines?

In theory, we could construct regression line for each school using the empirical Bayes estimates of the random effects. However, displaying 96 regression lines in a plot is probably a tad overwhelming. Instead, we will take advantage of the normality of the Bayes estimates to create confidence bands around the average regression line.

The variance of the intercepts equals 8.9467, resulting in a standard deviation of 2.9911. Normality implies that approximately 97% of the regression lines lie within two standard deviations of the average line. We can create these boundaries and display them in a plot with the predicted values. We use a multiplot to prevent cluttering a single plot with nine lines.

The input is:

```

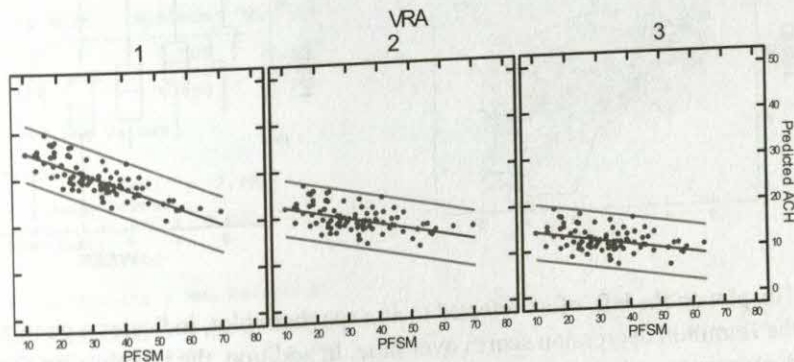
USE RESIDUALS1
LET UPPER=PRED0+5.9822
LET LOWER=PRED0-5.9822
LET VRA=1
IF FXD2(1) <> 1 AND FXD2(2)=1 THEN LET VRA=2
IF FXD2(1)=-1 AND FXD2(2)=-1 THEN LET VRA=3

BEGIN
PLOT UPPER*PFSM / MULTIPLY GROUP=VRA SMOOTH=LINEAR,
  SHORT YLAB='' COLOR=RED SIZE=0,
  FILL=1,0,0 LEGEND=NONE YMIN=0 YMAX=50,
  FTITLE=OFF
PLOT LOWER*PFSM / MULTIPLY GROUP=VRA SMOOTH=LINEAR,
  SHORT YLAB='' COLOR=RED SIZE=0,
  FILL=1,0,0 LEGEND=NONE YMIN=0 YMAX=50,
  FTITLE=OFF
PLOT PRED0*PFSM / MULTIPLY GROUP=VRA SMOOTH=LINEAR,
  SHORT YLAB='' COLOR=BLUE SIZE=0,
  FILL=1,0,0 LEGEND=NONE YMIN=0 YMAX=50,
  FTITLE=OFF
PLOT PRED1*PFSM / MULTIPLY GROUP=VRA FILL=1,0,0,
  LEGEND=NONE YMIN=0 YMAX=50,
  YLAB='Predicted ACH'

END

```

The output is:



Example 3

Longitudinal Data in Mixed Regression

Riesby et al. (1977) studied the relationship between desipramine and imipramine levels in plasma in 66 depressed patients classified as either endogenous or

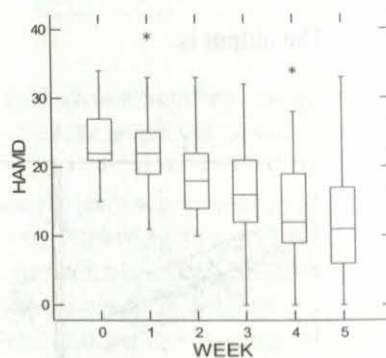
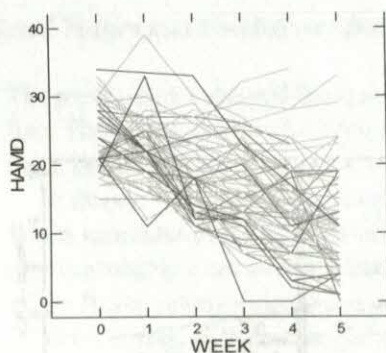
nonendogenous. After receiving a placebo for one week, the researchers administered a dose of imipramine each day for four weeks, recording the imipramine and desipramine levels at the end of each week. At the beginning of the placebo week and at the end of each week (including the placebo week), patients received a score on the Hamilton depression rating scale. Did the depression score change over time differently for each group of patients (endogenous vs nonendogeneous)?

A plot of the raw data often reveals general trends before fitting a model.

The input is:

```
MIX
USE RIESBY
BEGIN
CATEGORY WEEK
DENSITY HAMD * WEEK / BOX TICK=INDENT LOC=5IN,0IN
PLOT HAMD*WEEK / OVERLAY GROUP=ID LINE SIZE= 0,
LEGEND=NONE TICK=INDENT LOC=-1IN,0IN
CATEGORY
END
```

The output is:



The plot on the left, often referred to as a spaghetti plot, indicates a general decline in the Hamilton depression scores over time. In addition, the boxplots on the right demonstrate an increase in the variance of the depression scores over time.

Time as a Linear Effect

One potential model for the Riesby data includes a different intercept for each patient, as well as a linear change in depression score over time.

The input is:

```
MIX
USE RIESBY
RESET
MODEL HAMD
IDENTIFIER ID
RANDOM INTERCEPT WEEK
SAVE RESIDUALS1 / RESID
PLENGTH SHORT
ESTIMATE
```

The output is:

```
Perform 10 EM iterations
2 random terms
0 fixed terms
```

Numbers of Observations

```
Level 2 observations : 66
Level 1 observations : 375
```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
HAMD	0.000	39.000	17.637	7.190
INTERCEPT	1.000	1.000	1.000	0.000
WEEK	0.000	5.000	2.480	1.683

Starting Values

```
Mean:
    23.603    -2.405
Variance Terms:
    35.400     0.000    17.700
Residual:
    35.400
```

Final Results - MML Estimates

```
EM Iterations : 10
Fisher Iterations : 4
Total Iterations : 14
Log Likelihood : -1109.519
```

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	23.577	0.546	43.217	0.000
WEEK	-2.377	0.209	-11.393	0.000

Residual Variance

Estimate	Standard Error	Z	p-value

12.217 1.107 11.036 0.000

Random-Effect Variance & Covariance Term(s)

Estimate

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	12.629	
2	WEEK	-1.421	2.079

Standard Error

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	3.467	
2	WEEK	1.026	0.504

2

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	3.643	
2	WEEK	-1.385	4.124

p-value

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	0.000	
2	WEEK	0.166	0.000

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Random-Effect Covariances Expressed as Correlations

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	1.000	
2	WEEK	-0.277	1.000

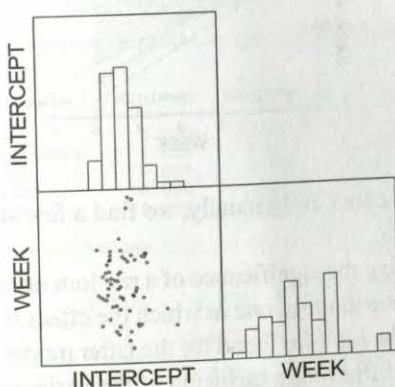
Correlation of the MML Estimates of the Fixed Terms

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	1.000	
2	WEEK	-0.449	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2	3	4
		VarCov1	VarCov2	VarCov3	Residual
1	VarCov1	1.000			
2	VarCov2	-0.590	1.000		
3	VarCov3	0.220	-0.588	1.000	
4	Residual	-0.180	0.169	-0.140	1.000

Empirical Bayes Estimates



Using the file of the residuals, we can display the overall average effect, as well as the individual trends.

The input is:

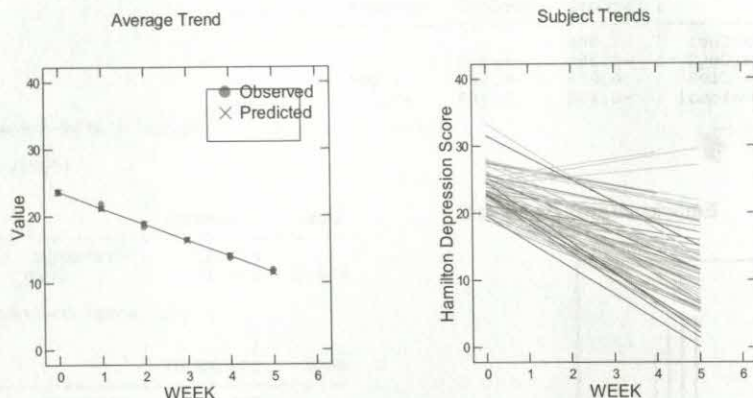
```

USE RESIDUALS1
BEGIN
DOT HAMD PRED1 *WEEK / OVERLAY TICK=INDENT TITLE='Average
Trend',
LEGEND=2.5IN,3.2IN
LLABEL='Observed','Predicted'
DRAW BOX / LOC=2.3IN,3IN HEIGHT=.7IN WIDTH=1.3IN
DOT PRED1*WEEK / LINE TICK=INDENT YLAB=''
PLOT PRED1*WEEK / OVERLAY GROUP=ID SMOOTH=LINEAR SHORT,
LEGEND=NONE SIZE=0 TICK=INDENT,
YLAB='Hamilton Depression Score',
TITLE='Subject Trends' LOC=6IN,0IN

```

END

The output is:



Overall, we see a general decline in scores. Individually, we find a few subjects who actually increased in depression score.

The preferred method of establishing the significance of a random effect involves a log-likelihood comparison between two models: one in which the effect is random and another in which the effect is fixed. The log-likelihood for the latter model (not shown) equals -1142.5944. The two models differ in the inclusion of the variance for *WEEK* and the covariance between *WEEK* and the intercept. Thus, according to the log-likelihood test, the difference between the log-likelihoods times -2 follows a chi-square distribution with two degrees of freedom. This value equals 66.15 and is significant. *WEEK* should be treated as a random effect.

Including Independent Variables

To a model with a linear effect of time, we can add the effect of diagnosis to look for differences due to this factor.

The input is:

```
MIX
USE RIESBY
RESET
MODEL HAMD = ENDOG WEEK*ENDOG
IDENTIFIER ID
RANDOM INTERCEPT WEEK
SAVE RESIDUALS1 / RESID
ESTIMATE
```

We include the interaction between *ENDOG* and *WEEK* to allow a separate trend for each type of diagnosis.

The output is:

Terms in the analysis and names of design matrix columns used for those terms
 WEEK * ENDOG
 FXD1
 Perform 10 EM iterations
 2 random terms
 2 fixed terms

Numbers of Observations

Level 2 observations : 66
 Level 1 observations : 375

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
HAMD	0.000	39.000	17.637	7.190
INTERCEPT	1.000	1.000	1.000	0.000
WEEK	0.000	5.000	2.480	1.683
ENDOG	0.000	1.000	0.547	0.498
FXD1	0.000	5.000	1.352	1.746

Starting Values

Mean: 22.518 -2.378
 Covariates: 1.974 -0.045
 Variance Terms: 34.721 0.000 17.361
 Residual: 34.721

Final Results - MML Estimates

EM Iterations : 10
 Fisher Iterations : 4
 Total Iterations : 14
 Log Likelihood : -1107.465

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	22.476	0.794	28.295	0.000
WEEK	-2.366	0.312	-7.587	0.000
ENDOG	1.988	1.069	1.860	0.063
FXD1	-0.027	0.419	-0.064	0.949

Residual Variance

Estimate	Standard Error	Z	p-value
12.218	1.107	11.037	0.000

Random-Effect Variance & Covariance Term(s)

Estimate

		1 INTERCEPT	2 WEEK
1	INTERCEPT	11.641	
2	WEEK	-1.402	2.077

Standard Error

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	3.296	
2	WEEK	1.003	0.504

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	3.531	
2	WEEK	-1.397	4.123

p-value

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	0.000	
2	WEEK	0.162	0.000

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Random-Effect Covariances Expressed as Correlations

		1	2
		INTERCEPT	WEEK
1	INTERCEPT	1.000	
2	WEEK	-0.285	1.000

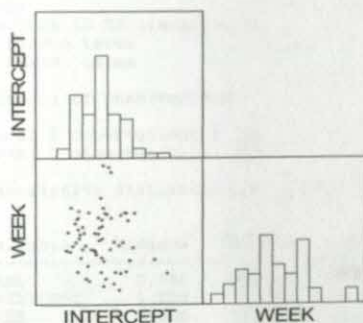
Correlation of the MML Estimates of the Fixed Terms

		1	2	3	4
		INTERCEPT	WEEK	ENDOG	FXD1
1	INTERCEPT	1.000			
2	WEEK	-0.451	1.000		
3	ENDOG	-0.743	0.335	1.000	
4	FXD1	0.335	-0.743	-0.457	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2	3	4
		VarCov1	VarCov2	VarCov3	Residual
1	VarCov1	1.000			
2	VarCov2	-0.601	1.000		
3	VarCov3	0.229	-0.598	1.000	
4	Residual	-0.189	0.173	-0.140	1.000

Empirical Bayes Estimates

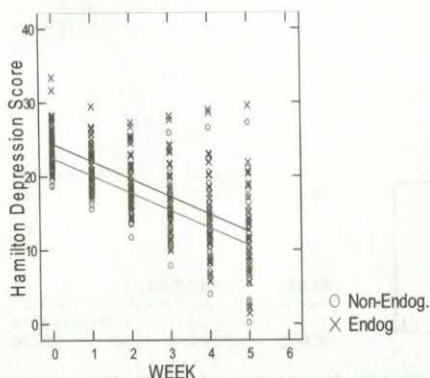


The parameter estimates suggest that *ENDOG* may have an effect ($p=.06$), but the two groups do not differ in their rate of change in depression ($p=.95$). We use scatterplots to display the results of this model.

The input is:

```
USE RESIDUALS1
BEGIN
PLOT PRED1*WEEK / OVERLAY GROUP=ENDOG,
    TICK=INDENT,
    YLAB='Hamilton Depression Score',
    LTITLE='' LLABEL='Non-Endog.', 'Endog',
    YMIN=0 YMAX=40
PLOT PRED0*WEEK / OVERLAY GROUP=ENDOG SMOOTH=LINEAR SHORT,
    SIZE=0 TICK=INDENT LEGEND=NONE,
    YLAB='' YMIN=0 YMAX=40
END
```


The output is:



The two lines are essentially parallel.

Time as a Quadratic Effect

In this example, we fit a quadratic effect of time to the Riesby data, ignoring the diagnosis of the patients.

The input is:

```
MIX
USE RIESBY
RESET
MODEL HAMD
IDENTIFIER ID
RANDOM INTERCEPT WEEK WEEK*WEEK
SAVE RESIDUALS1 / RESID
ESTIMATE
```

The output is:

Terms in the analysis and names of design matrix columns used for those terms:

WEEK * WEEK
 RND2
 Perform 10 EM iterations
 3 random terms
 0 fixed terms

Numbers of Observations

Level 2 observations : 66
 Level 1 observations : 375

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
HAMD	0.000	39.000	17.637	7.190
INTERCEPT	1.000	1.000	1.000	0.000
WEEK	0.000	5.000	2.480	1.683
RND2	0.000	25.000	8.976	8.734

Starting Values

Mean:
 23.759 -2.636 0.046
 Variance Terms:
 35.482 0.000 17.741 0.000 0.000 17.741
 Residual:
 35.482

Final Results - MML Estimates

EM Iterations : 10
 Fisher Iterations : 5
 Total Iterations : 15
 Log Likelihood : -1103.824

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	23.760	0.552	43.039	0.000
WEEK	-2.633	0.479	-5.496	0.000
RND2	0.051	0.088	0.583	0.560

Residual Variance

Estimate	Standard Error	Z	p-value
10.516	1.101	9.548	0.000

Random-Effect Variance & Covariance Term(s)

Estimate

	1	2	3
	INTERCEPT	WEEK	RND2
1 INTERCEPT	10.440		
2 WEEK	-0.915	6.638	
3 RND2	-0.112	-0.936	0.194

Standard Error

		1	2	3
		INTERCEPT	WEEK	RND2
1	INTERCEPT	3.579		
2	WEEK	2.418	2.746	
3	RND2	0.421	0.484	0.094

Z

		1	2	3
		INTERCEPT	WEEK	RND2
1	INTERCEPT	2.917		
2	WEEK	-0.379	2.418	
3	RND2	-0.266	-1.933	2.063

p-value

		1	2	3
		INTERCEPT	WEEK	RND2
1	INTERCEPT	0.002		
2	WEEK	0.705	0.008	
3	RND2	0.790	0.053	0.020

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Random-Effect Covariances Expressed as Correlations

		1	2	3
		INTERCEPT	WEEK	RND2
1	INTERCEPT	1.000		
2	WEEK	-0.110	1.000	
3	RND2	-0.079	-0.826	1.000

Correlation of the MML Estimates of the Fixed Terms

		1	2	3
		INTERCEPT	WEEK	RND2
1	INTERCEPT	1.000		
2	WEEK	-0.453	1.000	
3	RND2	0.299	-0.902	1.000

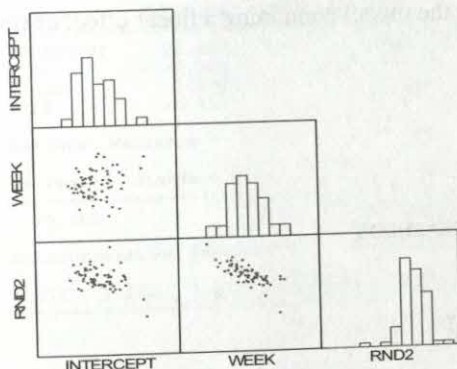
Correlation of the MML Estimates of Variance-Related Terms

		1	2	3	4	5	6
		VarCov1	VarCov2	VarCov3	VarCov4	VarCov5	VarCov6
1	VarCov1	1.000					
2	VarCov2	-0.603	1.000				
3	VarCov3	0.255	-0.612	1.000			
4	VarCov4	0.428	-0.909	0.579	1.000		
5	VarCov5	-0.201	0.518	-0.952	-0.544	1.000	
6	VarCov6	0.158	-0.401	0.833	0.445	-0.953	1.000
7	Residual	-0.263	0.280	-0.305	-0.244	0.320	-0.333

Correlation of the MML Estimates of Variance-Related Terms (contd...)

		7
		Residual
1	VarCov1	
2	VarCov2	
3	VarCov3	
4	VarCov4	
5	VarCov5	
6	VarCov6	
7	Residual	1.000

Empirical Bayes Estimates

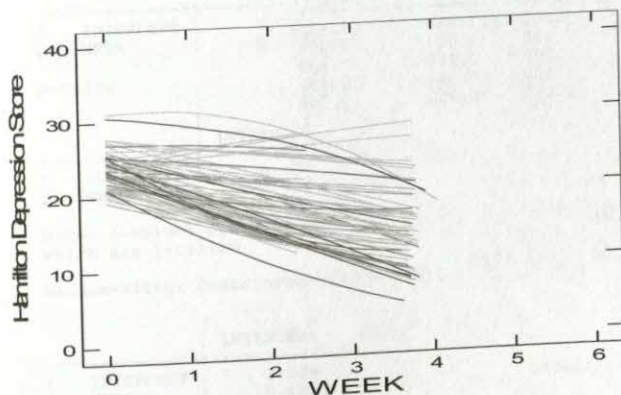


Although dividing the parameter estimate for the quadratic effect by its standard error suggests that a quadratic effect is not needed, the log-likelihood ratio test suggests otherwise ($-2 \cdot \Delta LL = 11.4$ on 4 degrees of freedom). The quadratic effects are plotted.

The input is:

```
USE RESIDUALS1
PLOT PRED1*WEEK / OVERLAY GROUP=ID SMOOTH=SPLINE SHORT,
LEGEND=NONE SIZE=0 TICK=INDENT,
YLAB='Hamilton Depression Score'
```

The output is:



Autocorrelated Errors

To illustrate the inclusion of autocorrelated errors in longitudinal data, we add a first-order autoregressive structure to the model containing a linear effect of time and an effect of diagnosis.

The input is:

```
MIX
USE RIESBY
RESET
MODEL HAMD = ENDOG WEEK*ENDOG
IDENTIFIER ID
RANDOM INTERCEPT WEEK
AUTO WEEK
SAVE RESIDUALS1 / RESID
ESTIMATE / NEM=0
```

The output is:

Terms in the analysis and names of design matrix columns used for those terms:

```
WEEK * ENDOG
FXD1
Perform 0 EM iterations
2 random terms
2 fixed terms
Autocorrelated Error Structure: AR(1)
```

Numbers of Observations

```
Level 2 observations : 66
Level 1 observations : 375
```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
HAMD	0.000	39.000	17.637	7.190
INTERCEPT	1.000	1.000	1.000	0.000
WEEK	0.000	5.000	2.480	1.683
ENDOG	0.000	1.000	0.547	0.498
FXD1	0.000	5.000	1.352	1.746

Starting Values

```
Mean:
    22.518    -2.378
Covariates:
    1.974    -0.045
Variance Terms:
    34.721    0.000    17.361
Residual:
    34.721
Auto Terms:
    0.200
```

Final Results - MML Estimates

EM Iterations : 0
 Fisher Iterations : 14
 Total Iterations : 14
 Log Likelihood : -1103.419

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	22.462	0.787	28.545	0.000
WEEK	-2.328	0.303	-7.688	0.000
ENDOG	1.870	1.060	1.764	0.078
FXD1	-0.016	0.408	-0.040	0.968

Residual Variance

Estimate	Standard Error	Z	p-value
15.489	1.925	8.046	0.000

Autocorrelation Term(s)

0.371	0.122	3.042	0.002
-------	-------	-------	-------

Random-Effect Variance & Covariance Term(s)

Estimate

		1 INTERCEPT	2 WEEK
1	INTERCEPT	3.901	
2	WEEK	0.340	1.276

Standard Error

		1 INTERCEPT	2 WEEK
1	INTERCEPT	5.307	
2	WEEK	1.264	0.580

Z

		1 INTERCEPT	2 WEEK
1	INTERCEPT	0.735	
2	WEEK	0.269	2.199

p-value

		1 INTERCEPT	2 WEEK
1	INTERCEPT	0.231	
2	WEEK	0.788	0.014

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Random-Effect Covariances Expressed as Correlations

		1 INTERCEPT	2 WEEK
1	INTERCEPT	1.000	
2	WEEK	0.152	1.000

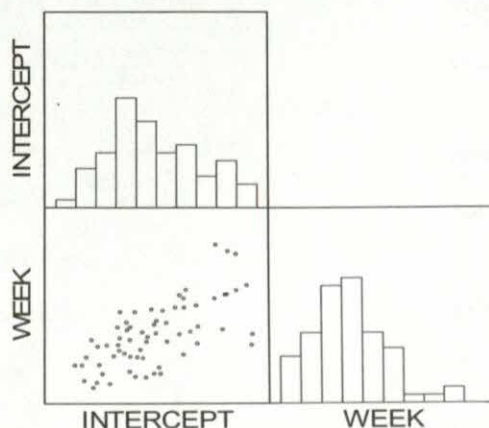
Correlation of the MML Estimates of the Fixed Terms

		1	2	3	4
		INTERCEPT	WEEK	ENDO	FXD1
1	INTERCEPT	1.000			
2	WEEK	-0.447	1.000		
3	ENDO	-0.742	0.332	1.000	
4	FXD1	0.332	-0.742	-0.454	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2	3	4	5
		VarCov1	VarCov2	VarCov3	Residual	AutoCor1
1	VarCov1	1.000				
2	VarCov2	-0.788	1.000			
3	VarCov3	0.564	-0.741	1.000		
4	Residual	-0.700	0.600	-0.530	1.000	
5	AutoCor1	-0.764	0.610	-0.539	0.685	1.000

Empirical Bayes Estimates



The log-likelihood test:

$$-2*(-1107.465+1103.419) = 8.1$$

has one degree of freedom due to the inclusion of the autoregressive parameter. This value is significant, suggesting the need for the autocorrelation structure. This structure has the following form:

1.00						
0.37	1.00					
0.14	0.37	1.00				
0.05	0.14	0.37	1.00			
0.02	0.05	0.14	0.37	1.00		
0.01	0.02	0.05	0.14	0.37	1.00	
0.002	0.01	0.02	0.05	0.14	0.37	1.00

Notice that although the inclusion of this matrix does not affect the fixed parameter estimates to any significant degree, the variances of the random parameters do change. The largest change occurs in the variance of the intercept, which drops from 11.6412 to 3.9009.

Example 4

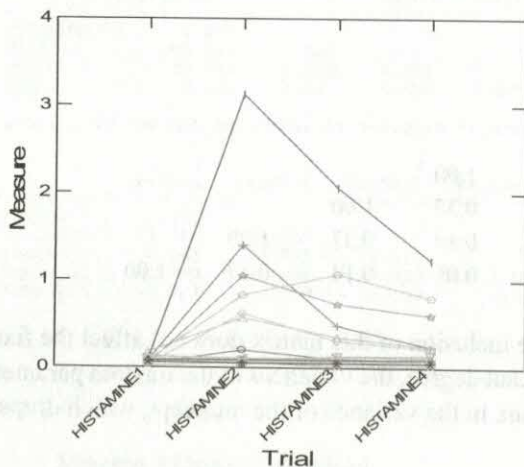
Multivariate Layout for Longitudinal Data

In this example, we analyze data having a multivariate layout from a study by Morrison and Zeppa (1963). In this study, mongrel dogs were divided into four groups of four. The groups received different drug treatments. The dependent variable, blood histamine in mg/mL, was then measured at four times after administration of the drug. The data are incomplete, since one of the dogs is missing the last measurement. We use a repeated-measures scatterplot to display.

The input is:

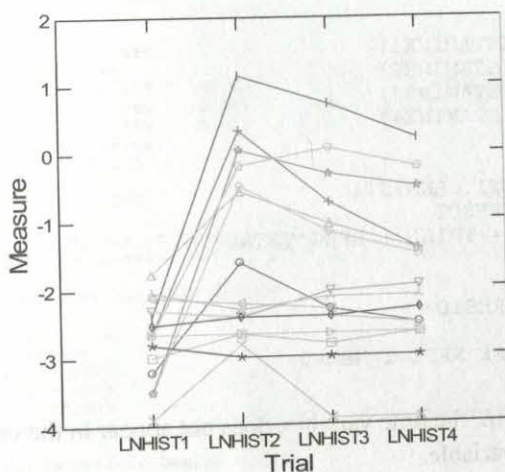
```
USE HISTAMINE
PLOT HISTAMINE1..HISTAMINE4 / OVERLAY REPEAT,
    GROUP=DOG,
    LINE LEGEND=NONE
```


The output is:



The variance in the histamine levels varies over time. In an effort to stabilize the variance, we apply a log-transformation.

```
USE HISTAMINE
let LN HIST1=LOG(HISTAMINE1)
let LN HIST2=LOG(HISTAMINE2)
let LN HIST3=LOG(HISTAMINE3)
let LN HIST4=LOG(HISTAMINE4)
PLOT LN HIST1..LN HIST4 / OVERLAY REPEAT GROUP=DOG,
    LINE LEGEND=NONE
```



The logged histamine levels now exhibit a similar spread of values at each measurement occasion. Subsequent analyses will use the logged values.

Random Intercept with Fixed Categorical Effects

To study the effects of the four drugs over time, we include the drug, a measure of time, and their interaction as fixed effects in a mixed regression model. To account for the dependencies due to taking repeated measurements on each dog, we include the dog as a random effect.

The dependent variable, histamine level, does not appear as a variable in the data file, *HISTAMINE*. Instead, the data file uses a multivariate layout, recording the histamine level across four variables representing time. To rearrange this data into a hierarchical structure, we use CONVERT.

The input is:

```

USE HISTAMINE
LET LN HIST1=LOG (HISTAMINE1)
LET LN HIST2=LOG (HISTAMINE2)
LET LN HIST3=LOG (HISTAMINE3)
LET LN HIST4=LOG (HISTAMINE4)
MIX
RESET
CONVERT HIST=LN HIST1..LN HIST4
CAT DRUG TRIAL / EFFECT
MODEL HIST = DRUG + TRIAL + DRUG*TRIAL
IDENTIFIER DOG
RANDOM INTERCEPT
SAVE RESIDUALS1 / RESID
PLENGTH SHORT
ESTIMATE / REPAR=OFF NREC=1 NEM=0

```

Notice that the variable *TRIAL*, the time variable, does not appear in the original data file. *CONVERT* creates this variable.

The output is:

Effects coding used for categorical variables in model
 Terms in the analysis and names of design matrix columns used for those terms:

```

DRUG
FXD1 (1) FXD1 (2) FXD1 (3)
TRIAL
FXD2 (1) FXD2 (2) FXD2 (3)
DRUG * TRIAL
FXD3 (1) FXD3 (2) FXD3 (3) FXD3 (4) FXD3 (5) FXD3 (6) FXD3 (7) FXD3 (8) FXD3 (9)
Perform 0 EM iterations
1 random terms
15 fixed terms

```

Numbers of Observations

```

Level 2 observations : 16
Level 1 observations : 63

```

Descriptive Statistics for all Variables

Variable	Minimum	Maximum	Mean	Standard Deviation
HIST	-3.912	1.141	-1.977	1.172
INTERCEPT	1.000	1.000	1.000	0.000
FXD1 (1)	-1.000	1.000	0.000	0.718
FXD1 (2)	-1.000	1.000	-0.016	0.707
FXD1 (3)	-1.000	1.000	0.000	0.718
FXD2 (1)	-1.000	1.000	0.016	0.707
FXD2 (2)	-1.000	1.000	0.016	0.707
FXD2 (3)	-1.000	1.000	0.016	0.707
FXD3 (1)	-1.000	1.000	0.000	0.508
FXD3 (2)	-1.000	1.000	0.000	0.508
FXD3 (3)	-1.000	1.000	0.000	0.508
FXD3 (4)	-1.000	1.000	0.016	0.492
FXD3 (5)	-1.000	1.000	0.016	0.492
FXD3 (6)	-1.000	1.000	0.016	0.492
FXD3 (7)	-1.000	1.000	0.000	0.508
FXD3 (8)	-1.000	1.000	0.000	0.508
FXD3 (9)	-1.000	1.000	0.000	0.508

Starting Values

Mean:

-1.984

Covariates:

-0.109	-0.487	1.091	-0.728	0.474	0.207
-0.069	0.458	-0.113	0.680	-0.486	-0.189
-1.399	0.551	0.512			

Variance Terms:

0.115

Residual:

0.577

Total Number of Level-2 Units = 16

Data for Level-2 Unit 1 which has 4 observations nested within

Dependent Variable Vector

	1
1	-3.219
2	-1.609
3	-2.303
4	-2.526

Random-effect Design Matrix

	1
1	1.000
2	1.000
3	1.000
4	1.000

Covariate Matrix

	1	2	3	4	5	6	7	8	9
1	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
2	1.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
3	1.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	1.000
4	1.000	0.000	0.000	-1.000	-1.000	-1.000	-1.000	-1.000	-1.000

Covariate Matrix (contd...)

	10	11	12	13	14	15
1	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	0.000	0.000	0.000
3	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000	0.000

Final Results - MML Estimates

EM Iterations	:	0
Fisher Iterations	:	5
Total Iterations	:	5
Log Likelihood	:	-23.789

Variable	Estimate	Standard Error	Z	p-value
INTERCEPT	-1.979	0.155	-12.731	0.000
FXD1 (1)	-0.115	0.269	-0.427	0.670
FXD1 (2)	-0.470	0.269	-1.745	0.081
FXD1 (3)	1.086	0.269	4.034	0.000
FXD2 (1)	-0.734	0.050	-14.626	0.000
FXD2 (2)	0.468	0.050	9.336	0.000
FXD2 (3)	0.201	0.050	4.015	0.000
FXD3 (1)	-0.063	0.087	-0.728	0.467
FXD3 (2)	0.463	0.087	5.350	0.000

FXD3(3)	-0.107	0.087	-1.237	0.216
FXD3(4)	0.664	0.088	7.569	0.000
FXD3(5)	-0.502	0.088	-5.731	0.000
FXD3(6)	-0.206	0.088	-2.349	0.019
FXD3(7)	-1.393	0.087	-16.084	0.000
FXD3(8)	0.556	0.087	6.422	0.000
FXD3(9)	0.518	0.087	5.980	0.000

Residual Variance

Estimate	Standard Error	Z	p-value
0.053	0.011	4.848	0.000

Random-Effect Variance & Covariance Term(s)

Estimate

	1
INTERCEPT	
1 INTERCEPT	0.373

Standard Error

	1
INTERCEPT	
1 INTERCEPT	0.137

Z

	1
INTERCEPT	
1 INTERCEPT	2.729

p-value

	1
INTERCEPT	
1 INTERCEPT	0.003

Note: p-values are 2-tailed except for those associated with variances, which are 1-tailed.

Calculation of the Intraclass Correlation

Residual Variance : 0.053
 Cluster Variance : 0.373
 Intraclass Correlation : $0.373 / (0.373 + 0.053) = 0.875$

Correlation of the MML Estimates of the Fixed Terms

	1	2	3	4	5	6
	INTERCEPT	FXD1(1)	FXD1(2)	FXD1(3)	FXD2(1)	FXD2(2)
1 INTERCEPT	1.000					
2 FXD1(1)	-0.001	1.000				
3 FXD1(2)	0.002	-0.334	1.000			
4 FXD1(3)	-0.001	-0.333	-0.334	1.000		
5 FXD2(1)	-0.003	0.002	-0.005	0.002	1.000	
6 FXD2(2)	-0.003	0.002	-0.005	0.002	-0.321	1.000
7 FXD2(3)	-0.003	0.002	-0.005	0.002	-0.321	-0.321
8 FXD3(1)	0.002	-0.001	0.003	-0.001	-0.005	-0.005
9 FXD3(2)	0.002	-0.001	0.003	-0.001	-0.005	-0.005
10 FXD3(3)	0.002	-0.001	0.003	-0.001	-0.005	-0.005
11 FXD3(4)	-0.005	0.003	-0.009	0.003	0.016	0.016
12 FXD3(5)	-0.005	0.003	-0.009	0.003	0.016	0.016
13 FXD3(6)	-0.005	0.003	-0.009	0.003	0.016	0.016

Mixed Regression

14	FXD3(7)	0.002	-0.001	0.003	-0.001	-0.005	-0.005
15	FXD3(8)	0.002	-0.001	0.003	-0.001	-0.005	-0.005
16	FXD3(9)	0.002	-0.001	0.003	-0.001	-0.005	-0.005

Correlation of the MML Estimates of the Fixed Terms (contd...)

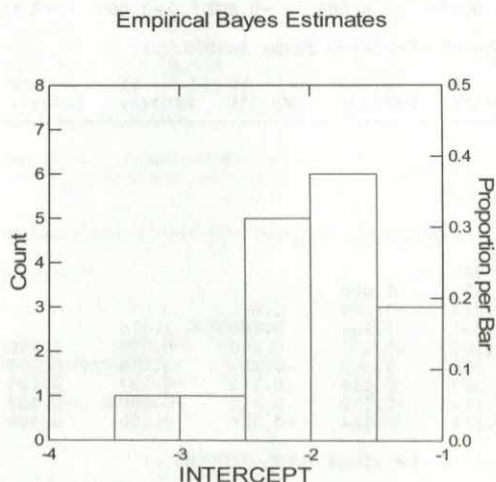
		7	8	9	10	11	12
		FXD2(3)	FXD3(1)	FXD3(2)	FXD3(3)	FXD3(4)	FXD3(5)
1	INTERCEPT						
2	FXD1(1)						
3	FXD1(2)						
4	FXD1(3)						
5	FXD2(1)						
6	FXD2(2)						
7	FXD2(3)	1.000					
8	FXD3(1)	-0.005	1.000				
9	FXD3(2)	-0.005	-0.329	1.000			
10	FXD3(3)	-0.005	-0.329	-0.329	1.000		
11	FXD3(4)	0.016	-0.337	0.100	0.100	1.000	
12	FXD3(5)	0.016	0.100	-0.337	0.100	-0.298	1.000
13	FXD3(6)	0.016	0.100	0.100	-0.337	-0.298	-0.298
14	FXD3(7)	-0.005	-0.329	0.114	0.114	-0.337	0.100
15	FXD3(8)	-0.005	0.114	-0.329	0.114	0.100	-0.337
16	FXD3(9)	-0.005	0.114	0.114	-0.329	0.100	0.100

Correlation of the MML Estimates of the Fixed Terms (contd...)

		13	14	15	16
		FXD3(6)	FXD3(7)	FXD3(8)	FXD3(9)
1	INTERCEPT				
2	FXD1(1)				
3	FXD1(2)				
4	FXD1(3)				
5	FXD2(1)				
6	FXD2(2)				
7	FXD2(3)				
8	FXD3(1)				
9	FXD3(2)				
10	FXD3(3)				
11	FXD3(4)				
12	FXD3(5)				
13	FXD3(6)	1.000			
14	FXD3(7)	0.100	1.000		
15	FXD3(8)	0.100	-0.329	1.000	
16	FXD3(9)	-0.337	-0.329	-0.329	1.000

Correlation of the MML Estimates of Variance-Related Terms

		1	2
		VarCov1	Residual
1	VarCov1	1.000	
2	Residual	-0.020	1.000



DRUG is the first fixed effect variable included in the model so the coded variable for *DRUG* uses a root name of *FXDI*. The variable has four levels so three effect-coded variables are needed. Furthermore, *TRIAL* has four levels so three effect-coded variables are needed. The final fixed effect, the interaction, involves the crossing of two variables, each having four levels, so $(4-1)*(4-1)$ effect-coded variables are needed.

The output includes a listing of the converted data for the first level 2 unit. The dependent variable vector displays the log transformed histamine levels. The random-effect design matrix shows that the dog in question is the first. The covariate matrix corresponds to the fixed effects in the model as follows:

- The first three columns represent *DRUG*. The first dog received drug 1, so the first column equals 1 and the next two equal 0.
- The next three columns represent *TRIAL*, comparing each measurement occasion to the last.
- The final nine columns represent the interaction and result from crossing the first three columns with the second three.

Looking at the parameter estimates for the fixed effects, we find that although some contrasts are not significant, each factor does exhibit a significant effect. This suggests that the histamine levels varied over time differently for each drug. The predicted values may shed some light on how these factors interact.

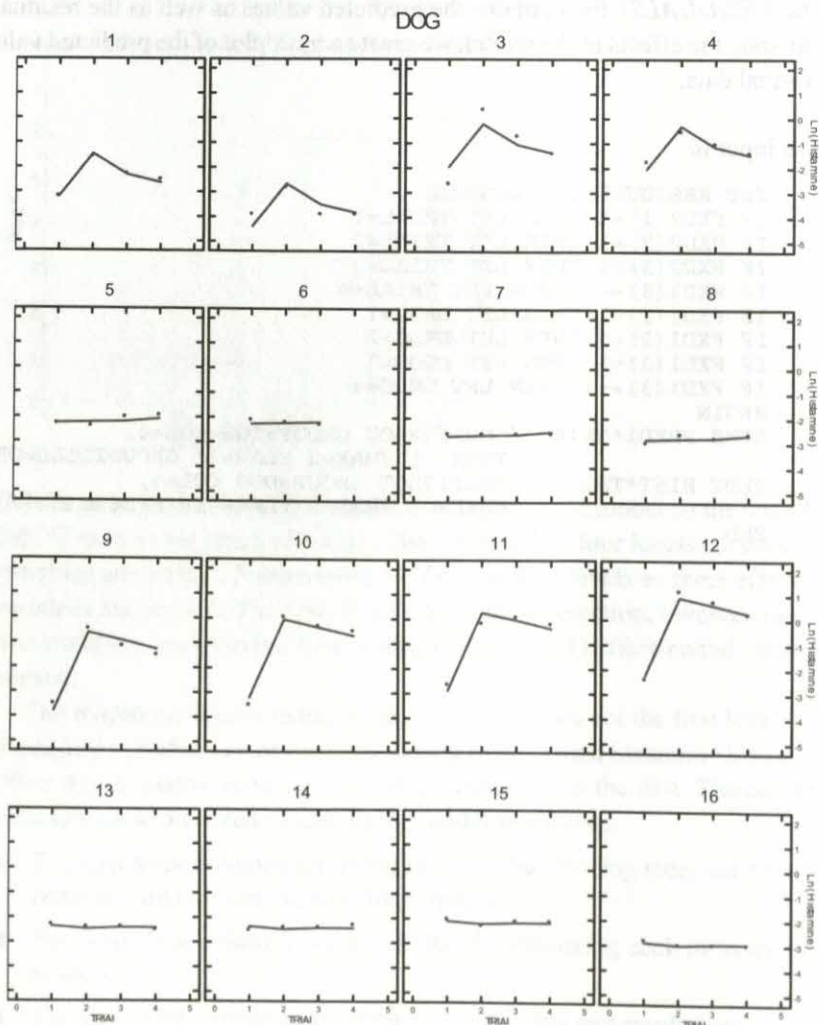
Predicted Values

The *RESIDUALS1* file contains the predicted values as well as the residuals. To illustrate the effects in the model, we create a multiplot of the predicted values with the original data.

The input is:

```
USE RESIDUALS1 / NONAMES
IF FXD2(1)=1 THEN LET TRIAL=1
IF FXD2(2)=1 THEN LET TRIAL=2
IF FXD2(3)=1 THEN LET TRIAL=3
IF FXD2(3)=-1 THEN LET TRIAL=4
IF FXD1(1)=1 THEN LET DRUG=1
IF FXD1(2)=1 THEN LET DRUG=2
IF FXD1(3)=1 THEN LET DRUG=3
IF FXD1(3)=-1 THEN LET DRUG=4
BEGIN
LINE PRED1*TRIAL / MULTIPLY GROUP=DOG COL=4,
                      YMIN=-5 YMAX=2 YLAB='' GROUPTITLE=OF
PLOT HIST*TRIAL / MULTIPLY GROUP=DOG COL=4,
                  YMIN=-5 YMAX=2 YLAB='Ln(Histamine)'
END
```


The output is:



The multiplot clearly illustrates the interaction between *DRUG* and *TRIAL*. Those dogs receiving drugs 2 and 4, the second and fourth rows of the multiplot, show very little change in histamine level over time. The remainder of the dogs exhibited a sharp increase in histamine level, followed by a general decrease.

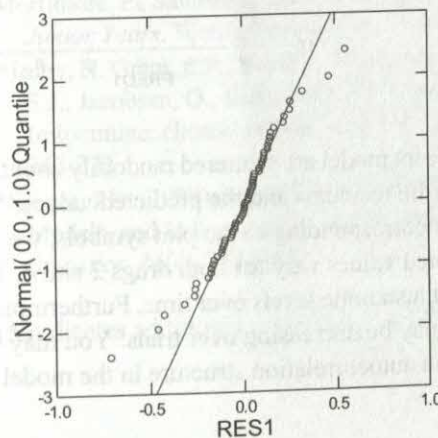
Residuals

To examine the adequacy of the model, we focus on the residuals. We can assess their normality using a probability plot.

The input is:

```
USE RESIDUALS1 / NONAMES
IF FXD2(1)=1 THEN LET TRIAL=1
IF FXD2(2)=1 THEN LET TRIAL=2
IF FXD2(3)=1 THEN LET TRIAL=3
IF FXD2(3)=-1 THEN LET TRIAL=4
IF FXD1(1)=1 THEN LET DRUG=1
IF FXD1(2)=1 THEN LET DRUG=2
IF FXD1(3)=1 THEN LET DRUG=3
IF FXD1(3)=-1 THEN LET DRUG=4
PLOT RES1 / NORMAL SMOOTH=MIDRANGE
```

The output is:

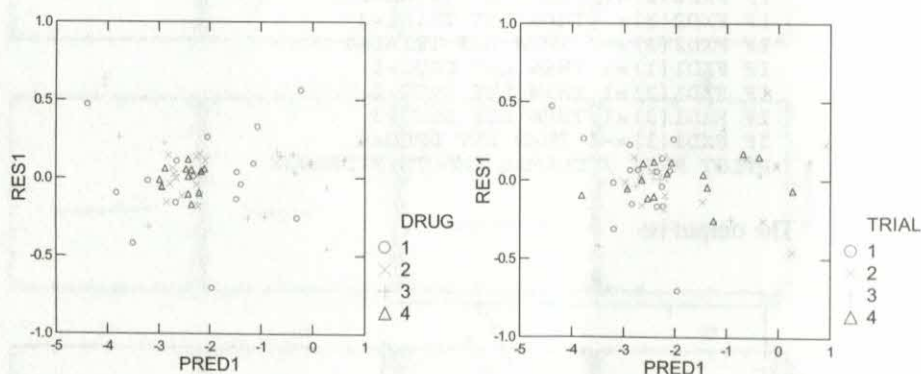


The straight line along which the residuals lie indicates that normality of the residuals appears to be satisfied. We can examine other model assumptions by plotting the residuals against the predicted values.

The input is:

```
BEGIN
PLOT RES1*PRED1 / OVERLAY GROUP=DRUG LOC=-1IN,0IN,
                  LEGEND=4.2IN,.1IN
PLOT RES1*PRED1 / OVERLAY GROUP=TRIAL LOC=5IN,0IN,
                  LEGEND=4.2IN,.1IN
END
```

The output is:



The residuals for the random intercept model are scattered randomly about zero. There appears to be no relation between the residuals and the predicted values.

If we focus on the fixed effects corresponding to the plot symbols, we see a very small range over which the predicted values vary for both drugs 2 and 4. These drugs both resulted in relatively constant histamine levels over time. Furthermore, it appears that the variance of the residuals may be decreasing over trials. You may want to examine the effects of including an autocorrelation structure in the model.

Computation

Algorithms

Mixed regression uses marginal maximum likelihood to estimate the parameters of the model. The procedure involves a combination of the EM algorithm and Fisher scoring. For details, see Hedeker and Gibbons (1996).

References

- *Andersen, A.H., Jensen, E.B., and Schou, G. (1981). Two-way analysis of variance with correlated errors. *International Statistical Review*, 49, 153-167.
- Bryk, A.S. and Raudenbush, S.W. (2001). *Hierarchical Linear Models*, 2nd ed. Sage: Newbury Park, CA.
- de Leeuw, J. and Kreft, I.G.G. (1986). Random Coefficient Models for Multilevel Analysis. *Journal of Educational Statistics*, 11, 57-85.
- Goldstein, H.(1987). *Multilevel models in educational and social research*. London: Griffin.
- Hedeker, D. and Gibbons, R.D. (1996). MIXREG: a computer program for mixed-effects regression analysis with autocorrelated errors. *Computer Methods and Programs in Biomedicine*, 49, 229-252.
- Longford, N. J. (1993). *Random Coefficient Models*. Clarendon Press: Oxford.
- Morrison, K.J. and Zeppa, R. (1963). Histamine-introduced hypotension due to morphine and arfonad in the dog, 3, 313-317. *Journal of Surgical Research*.
- Mortimore, P., Sammons, P., Stoll, L., Lewis, D., and Ecob, R. (1988). *School Matters, the Junior Years*. Wells: Open Books.
- Riesby, N. Gram, L.F., Bech, P., Nagy, A., Peterson, G.O., Ortmann, J., Ibsen, I., Dencker, S.J., Jacobsen, O., Krautwald, O., Sondergaard, I., and Christiansen, J. (1977). Imipramine: clinical effects and pharmacokinetic variability. *Psychopharmacology*, 54, 263-272.
- *Singer, J.D. (1998). Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. *Journal of Educational and Behavioral Statistics*, 24(4), 323-355.

(* indicates additional reference.)

Acronym & Abbreviation Expansions

A

ABS - absolute value
ACF - autocorrelation function
ACOLOR - color axes
ACS - arccosine
ACT - actuarial life table
AD test - Anderson Darling test
ADDTREE - additive trees
ADFG - asymptotically distribution free estimate biased, Gramian
ADFU - asymptotically distribution free estimate unbiased
ADJSEASON - seasonal adjustment
AHMAX - maximum extent
AHMIN - minimum extent
AIC - Akaike information criterion
AID - automatic interaction detection
ALT - alternative
ANCOVA - analysis of covariance
ANG1 - deviation of angles from north in a clockwise direction
ANG2 - deviation of angles from horizontal (for 3D models)
ANG3 - tilt angle
ANOVA - analysis of variance
ANOVAHYPO - hypothesis tests in analysis of variance
AR - autoregressive
ARIMA - autoregressive integrated moving average
ARL - average run length

ARMA - autoregressive moving average
ARS - adaptive rejection sampling
ASCII - American Standard Code for Information Interchange
ASE - asymptotic standard error
ASN - arcsine
ATH - arc hyperbolic tangent
ATN - arctangent
AVERT - vertical extent
AVG - average

B

BC - Bray-Curtis similarity measure
BCa - Bias Corrected and accelerated
BCF - Beta cumulative function
BDF - Beta density function
BETACORR - beta correction
BIC - Bayesian information criterion
BIF - Beta inverse function
BMP - Windows bitmap
BOF - beginning-of-file
BOG - beginning-of-BY group
BONF - Bonferroni
BOOT - bootstrap
BRN - Beta random number

C

CART - classification and regression trees
CBSTAT - column basic statistics
CCF - Cauchy cumulative function
CCF - cross-correlation function
CDF - Cauchy density function
cdf/CF - cumulative distribution function
CDFUNC - coefficients for canonical variables

CFUNC - coefficients for the classification functions
 CGM - Computer graphics metafile: binary or clear text
 CHAZ - cumulative hazard
 CHISQ - Chi-square distribution
 CHOL - Cholesky decomposition
 CI - confidence interval
 CIF - Cauchy inverse function
 CIM - confidence interval of mean
 CLASS - classification
 CLSTEM - stem and leaf plot for column
 CMeans - canonical scores of group means
 CMULTIVAR - multiple string variables
 COEF - coefficients
 COL/col - column
 COLPCT - Column percentages
 CONFIG - configuration
 CONT - Contingency coefficient
 CONV - convergence
 CORAN - correspondence analysis
 CORR - correlations
 CORR1 - single correlation coefficient
 CORR2 - equality of two correlations
 COV - covariance
 Cp - process capability index
 CPL - process capability based on lower specification limit
 CPU - process capability based on upper specification limit
 Cpk-Process capability index for off-centered process
 CR - confidence region
 CRA - cost of response above UTL
 CRB - cost of response below LTL
 CRN - Cauchy random number
 CSCORE - canonical scores
 CSIZE - size of characters
 CSQ - Chi-square
 CSTATISTICS - column statistics
 CSV - comma separated values

CUSUM - cumulative sum
 CUSUM HI - Upper cumulative sum
 CUSUM LO - Lower cumulative sum
 CV - coefficient of variation
 CVI - cross validation index

D

DBF - Dbase files
 DC - deciles of risk
 DECF - Double exponential cumulative function
 DEDF - Double exponential density function
 DEIF - Double exponential inverse function
 DENFUN - density function
 dep. - dependent
 DERN - Double exponential random number
 DET - determinant
 DEVI - deviates (observed values - expected values)
 DEXP - Double exponential distribution
 df - degrees of freedom
 DF - distribution function
 DHAT - estimated distance
 DIF - data interchange format
 DIM - dimension
 DISCRIM - discriminant analysis
 DIST - distance
 DIT - dot histogram
 DOE - design of experiments
 DOS - disc operating system
 DPMO - defects per million opportunities
 DPU - defects per unit
 DTA - Stata files
 DUCF - Discrete uniform cumulative function
 DUDF - Discrete uniform density function
 DUIF - Discrete uniform inverse function
 DUNIFORM - Discrete uniform
 DURN - Discrete uniform random number
 DWLS - distance weighted least-squares

E

ECF - Exponential cumulative function

EDF - Exponential density function
 EEXP - extreme value exponential
 EIF - Exponential inverse function
 EIGEN - eigenvalues
 ELAMBDA - $\exp(\lambda)$
 EM - expectation-maximization
 EMF - Windows enhanced metafile
 ENCF - Logit normal cumulative function
 ENDF - Logit normal density function
 ENIF - Logit normal inverse function
 ENORMAL - Logit normal
 ENRN - Logit normal random number
 EOF - end-of-file
 EOG - end-of-BY group
 EPS - Encapsulated postscript
 ERN - Exponential random number
 ES - exhaustive search
 ESS - error sum of squares
 EW - extreme value Weibull
 EWMA - exponentially weighted moving average
 EXP/exp - exponential/ expected

F

FAR - false-alarm rates
 FCF - F cumulative function
 FCOLOR - color foreground
 FDF - F density function
 FIF - F inverse function
 FINV - inverse of the F cumulative
 FITC - fitting distribution: continuous
 FITD - fitting distribution: discrete
 FITDIST - fitting distributions
 Flexibeta - flexible beta
 FPLOTT - function plots
 FRN - F random number
 FTD - folded trellis detector
 FTDEV - Freeman-Tukey deviate
 FULLCOND - full conditional
 FUN - function

G

GCF - Gamma cumulative function
 GCOR - groupwise correlation matrix
 GCOV - groupwise covariance matrix
 GCV - generalized cross validation
 GDF - Gamma density function
 GECF - Geometric cumulative function
 GEDF - Geometric density function
 GEIF - Geometric inverse function
 GEN - general Toeplitz structure
 GERN - Geometric random number
 GG - Greenhouse Geisser
 GIF - Gamma inverse function
 GIF - Graphics Interchange Format
 GLM - generalized linear models
 GLMHYPOT - hypothesis tests in general linear model
 GLMPOST - post hoc estimate for repeated measures in general linear model
 GLS - generalized least-squares
 GMA - geometric moving average
 GN - Gauss-Newton method
 GOCF - Gompertz cumulative function
 GODF - Gompertz density function
 GOIF - Gompertz inverse function
 GORN - Gompertz random number
 GRN - Gamma random number
 GUCF - Gumbell cumulative function
 GUDF - Gumbell density function
 GUIF - Gumbell inverse function
 GURN - Gumbell random number

H

H & L - Hosmer and Lemeshow
 HC - heteroscedasticity-consistent
 HCF - Hypergeometric cumulative function
 HDF - Hypergeometric density function
 HF - Huynh-Feldt
 HGEOMETRIC - hypergeometric
 HIF - Hypergeometric inverse function
 HIST - histogram
 HKB - Hoerl, Kennard, and Baldwin

H-L trace - Holding-Lawley trace

HR - hit-rates

HRN - Hypergeometric random number

HSD - honestly significant differences

HTERM - terms tested hierarchically

HTML - hyper text markup language

HYMH - hybrid Metropolis-Hastings

I

IF - Inverse cumulative distribution function

IGAUSSIAN - inverse Gaussian

IGCF - Inverse Gaussian cumulative function

IGDF - Inverse Gaussian density function

IGIF - Inverse Gaussian inverse function

IGRN - Inverse Gaussian random number

IIDMC - independently and identically distributed Monte Carlo

IMPSAMPI - importance sampling integration

IMPSAMPR - importance sampling ratio

I-MR - individual and moving range

Ind/indep - independent

IndMH - Independent Metropolis-Hastings

INDSCAL - individual differences scaling

INITSAMP - initial sample

INTEG FUN - integrated function

IPA - iterated principal axis

ITER - iterations

J

JACK - jackknife

JCLASS - jackknifed classification

JMP - JMP v3.2 data files

JPEG/JPG - joint photographic experts group

K

K-M - Kaplan-Meier

KNBD - kth nearest neighborhood

KRON - Kronecker product

K-S test - Kolmogorov-Smirnov test

KS1 - one sample Kolmogorov-Smirnov tests

KS2 - two sample Kolmogorov-Smirnov tests

L

LAD - least absolute deviations

LB - larger the better

LCF - Logistic cumulative function

LCHAZ - log cumulative hazard

LCL - lower control limit

LCONV - log-likelihood convergence criteria

LDF - Logistic density function

LGM - log gamma

LGST - logistic

LIF - Logistic inverse function

L-L/LL - log likelihood

LMS - least median of squares

LMSREG - least median of squares regression

LNCF - Lognormal cumulative function

LNDF - Lognormal density function

LNIF - Lognormal inverse function

LNOR/LNORM - lognormal

LNRN - Lognormal random number

loc - location

LOG1 - one-parameter logistic (Rasch)

LOG2 - two-parameter logistic

LOGIT - logistic regression

LOGITHYPO - hypothesis tests in logistic regression

LOGLIN - loglinear modeling

LR - likelihood ratio

LRCHI - likelihood ratio chi-square

LRDEV - likelihood ratio of deviate

LRN - Logistic random number

LS - least-squares

LSD - least significant difference

LSL - lower specification limit

LSQ - least-squares

LTAB - life tables

LTL - lower tolerance limit

LW - Lawless and Wang

M

MA - moving average

- MAD - mean absolute deviation
 MAHAL - Mahalanobis distances
 MANCOVA - multivariate analysis of covariance
 MANOVA - multivariate analysis of variance
 MANOVAHYPO - hypothesis tests in MANOVA
 MANOVAPOST - post hoc estimate for repeated measures in MANOVA
 MAR - missing at random
 MAX - maximum
 MAXSTEP - maximum number of steps
 MCAR - missing completely at random
 MCMC - Markov Chain Monte Carlo
 MDPREF - multidimensional preference
 MDS - multidimensional scaling
 MIN - minimum
 M-H- Metropolis-Hastings
 MIS - number of missing values
 MIX - mixed regression
 MIXHIER - mixed regression for data having a hierarchical structure
 MIXMULTY - mixed regression for data having a multivariate structure
 ML - Maximum Likelihood
 MLA - maximum likelihood analysis
 MLE - maximum likelihood estimate
 MML - maximum marginal likelihood
 MRC - Multiple Regression and Correlation
 MS - mean squares
 MSE - mean square error
 MSIGMA - sigma measurement
 MT - Mersenne-Twister
 MTW - MINITAB v11 data files
 MU2 - Guttman's mu2 monotonicity coefficients
 MULTIVAR - multiple variables
 MW - minimum within sum of squares deviations
 MWL - maximum Wishart likelihood

 N
 NAR - non-stationary first-order autoregressive
 NB - nominal the best
 NBB - nominal-the-best: bilateral tolerance
 NBCF - Negative binomial cumulative function
 NBD - number of active bounds on parameter values
 NBDF - Negative binomial density function
 NBIF - Negative binomial inverse function
 NBINOMIAL - Negative binomial
 NBRN - Negative binomial random number
 NBU - nominal-the-best: unilateral tolerance
 NCAT - number of categories
 NCF - Binomial cumulative function
 NCOL - number of columns
 NDF - Binomial density function
 NDMAX - maximum number of points
 NDMIN - minimum number of points
 NEM - number of EM iterations
 NEXPO - negative exponential
 NIF - Binomial inverse function
 NIPALS - Nonlinear iterative partial least Squares
 NLAG - number of lags
 NLLOSS - nonlinear loss functions
 NLMODEL - nonlinear models
 NMIN - minimum count
 NMULTIVAR - multiple numeric variables
 NONLIN - nonlinear models
 NP-Number nonconforming
 NPAR - nonparametric
 NREC - non-recreationist
 NRN - Binomial random number
 NROW - number of rows
 NRP - number of apparently redundant parameters
 NSAMP - number of sub-samples
 NSPLIT - maximum number of splits
 NX - number of nodes along the x axis
 NXDIS - number of discretization points in the x (North) direction
 NY - number of nodes along the y axis
 NYDIS - number of discretization points in the y (East) direction
 NZ - number of nodes along the z axis

NZDIS - number of discretization points in the z (Depth) direction

O

Obs-observed

OBSFREQ - observed frequency

OC - operating characteristic

ODBC - open database capture and connectivity

OFREQ - outlier frequencies

OLS - ordinary least-squares

ORTHEQ-Equally Spaced Orthogonal component

ORTHUN- Unequally Spaced Orthogonal component

P

P - Proportion nonconforming

PACF - Pareto cumulative function

PACF - partial autocorrelation function

PADF - Pareto density function

PAIF - Pareto inverse function

PARAM - parameters

PARN - Pareto random number

PCA - process capability analysis

PCF - iterated principal axis factoring

PCF - Poisson cumulative function

PCNTCHANGE - percentage change

PCT - Macintosh PICT

PDF - Poisson density function

pdf - probability density function

PDL - polynomial distributed lag

PERMAP - perceptual mapping

PIF - Poisson inverse function

PLIMITS - probability limits

PLS - partial least squares

pmf - probability mass function

PMIN - minimum proportion

PNG - Portable Network Graphics

POLY - polygon

POSAC - partially ordered scalogram analysis with coordinates

P-P - probability plot

PP - process performance

Ppk - Process performance index for off-centered process

PPL - process performance based on lower specification limit

PPM - parts per million

PPU - process performance based on upper specification limit

PRE - percentage reduction error

PREFMAP - preference mapping

PRN - Poisson random number

PROB - probability

PROP1 - single proportion

PROP2 - equality of two proportions

PS - PostScript

PVAF/p.v.a.f. -- present value annuity factor

p-value - probability value

Q

QC - quality control

QMLE - quasi maximum likelihood estimate

QNTL - quantiles

QPLOT - quantile plots

Q-QPLOT - two sample quantile plot

QRD - QR decomposition

QS - quick search

QSK - quantitative symmetric similarity coefficients (or Kulczynski measure)

QUASI - Quasi-Newton method

R

R & R - repeatability and reproducibility

R chart - range chart

RADMAX - maximum horizontal direction for the search radius

RADMIN - minimum horizontal direction for the search radius

RAND - random

RANDSAMP - random sampling

RANKREG - rank regression

- RBSTAT - row basic statistics
 RCF - Rayleigh cumulative function
 RDF - Rayleigh density function
 RDISCRIM - robust discriminant
 RDIST - robust distance
 RDVER - vertical direction for the search radius
 REPAR - reparametrize
 REPS - replicates
 RESID - residuals
 RIF - Rayleigh inverse function
 RJS - rejection sampling
 RMS - root mean square
 RMSEA - root mean square error of approximation
 RMSSTD - root mean square standard deviation
 ROC - receiver operating characteristic
 ROWPCT - Row percentages
 RRN - Rayleigh random number
 RS - response surface
 RSE - robust standard errors
 RSEED - random seed
 RSM - response surface methods
 RSQ - stress and squared correlation
 RSS - residual sum of squares
 RSTATISTICS - row statistics
 RTF - rich text format
 RWM-H - random walk Metropolis-Hastings
 RWSTEM - stem and leaf plot for rows
- S
- S chart - standard deviation control chart
 SANG1 - angle (in degrees) of the first minor axis of the search ellipsoid
 SANG2 - angle (in degrees) of the major axis of the search ellipsoid
 SANG3 - angle (in degrees) of the second minor axis of the search ellipsoid
 SAV - SPSS files
 SB - smaller the better
 sc - scale
 SC - set correlation
- SCDFUNC - standardized coefficients for canonical variables
 SCF - Studentized cumulative function
 SD - standard deviations
 sd2/sas7bdat - SAS v9 files
 SDF - Studentized density function
 SE/se/S.E. - standard error
 SEK - standard error of kurtosis
 SEM - standard error of mean
 SES - standard error of skewness
 shp - shape
 SIF - Studentized inverse function
 SIMPLS - Straight-forward Implementation of Partial Least Squares
 SKMEAN - simple kriging mean
 SL - specification limit
 SMIN - minimum split value
 SPLOM - scatter plot matrix
 SQL - structured query language
 SQRT/SQR - square-root
 SRN - Studentized random number
 SRWR - sum of rank weighted residuals
 SS - sum of squares
 SSCP - sum of squares and cross products
 STA - Statistica v5 data files
 STAND - standardized deviates
 SVD - singular value decomposition
 SW - Shapiro-Wilks
 SYC/CMD - SYSTAT command Files
 SYZ/SYD/SYS - SYSTAT data files
 SYO - SYSTAT output files
- T
- T1 - one-sample t-test
 T2 - two-sample t-test
 TANALYZE - Taguchi design: analyze
 TCF - t cumulative function
 TCOR - total correlation
 TCOV - total covariance
 TDF - t density function
 TESTAT - Test Item Analysis

- TESTATCL - classical test item analysis
 TESTATLOG - logistic item response analysis
 TETRA - tetrachoric correlations
 TGENERATE - Taguchi design: generate
 TIF - t inverse function
 TIFF - Tagged Image File Format
 TLOG - log time
 TLOSS - Taguchi's Loss Function
 TNH - hyperbolic tangent
 TOHC0 - Hypothesis Testing: Zero correlation
 TOHC1 - Hypothesis Testing: Specific correlation
 TOHC2 - Hypothesis Testing: Equality of two correlation coefficients
 TOHP1 - Hypothesis Testing: Single proportion
 TOHP2 - Hypothesis Testing: Equality of two proportions
 TOHT1 - Hypothesis Testing: One sample t-test
 TOHT2 - Hypothesis Testing: Two sample t-test
 TOHTPAIRED - Hypothesis Testing: Paired t-test
 TOHV1 - Hypothesis Testing: Single variance
 TOHV2 - Hypothesis Testing: Two variances
 TOHVN - Hypothesis Testing: Several variances
 TOHZ1 - Hypothesis Testing: One sample z-test
 TOHZ2 - Hypothesis Testing: Two sample z-test
 TOL - tolerance
 TPLLOT - time series plot
 TPREDICT - Taguchi design: predict
 TRCF - Triangular cumulative function
 TRDF - Triangular density function
 TRI - triangular
 TRIF - Triangular inverse function
 TRIM - trimmed mean
 TRN - t random number
 TRP - transpose
 TRRN - Triangular random number
 TSFOURIER - Fourier decomposition of time series
 TSIV - Two-Stage Instrumental Variables
 TSLS - Two-Stage Least Squares
 TSP - traveling salesman path
 TSQ chart - Hotelling's T^2 chart
 TSSMOOTH - smoothing time series
 TXT - text format
- U
- U chart - chart showing defects per unit
 UCF - Uniform cumulative function
 UCL - upper control limit
 UDF - Uniform density function
 UIF - Uniform inverse function
 UNCE - uncertainty coefficient
 URN - Uniform random number
 USL - upper specification limit
 UTL - upper tolerance limit
- V
- VAR - variance
 VIF - variance inflation factor
- W
- WB - Weibull
 WCF - Weibull cumulative function
 WCOR - pooled within-group correlation
 WCOV - pooled within-group covariance
 WDF - Weibull density function
 WHISKER - Box-and-Whisker plot
 WIF - Weibull inverse function
 WMF - Windows metafile
 WRN - Weibull random number
- X
- XCF - Chi-square cumulative function
 XDF - Chi-square density function
 XIF - Chi-square inverse function
 XLAG - separation distance between lags
 XLS - excel format
 XLTOL - tolerance for lags
 XMAX - maximum along x axis
 XMIN - minimum along x axis

X-MR chart - Individuals and moving range chart
 XPT/TPT - SAS transport files
 XRN - Chi-square random number
 XTAB - Crosstabulations

Y

YMAX - maximum along y axis
 YMIN - minimum along y axis

Z

Z1 - one-sample z-test
 Z2 - two-sample z-test
 ZCF - Normal cumulative function
 ZDF - Normal density function
 ZICF - Zipf cumulative function
 ZIDF - Zipf density function
 ZIF - Normal inverse function
 ZIIF - Zipf inverse function
 ZIRN - Zipf random number
 ZMAX - maximum along z axis
 ZMIN - minimum along z axis
 ZRN - Normal random number

Index

A

A matrix, II-192

accelerated failure time distribution, IV-433

ACF plots, IV-529

additive trees, I-80, I-91

AIC and Schwarz's BIC, II-39, II-108, II-292, II-300, II-344, II-385, III-1, III-258, IV-99, IV-427
see linear models, II-17

Akaike Information Criterion, III-458

alpha level, IV-22, IV-28

alternative hypothesis, I-13, IV-20

analysis of covariance, II-153, II-209
examples, II-170

analysis of variance, II-107

AIC and Schwarz's BIC, II-108

algorithms, II-171

assumptions, II-25

between-group differences, II-32

commands, II-121

compared to loglinear modeling, III-95

compared to regression trees, I-45

contrasts, II-28, II-113, II-115, II-116

data format, II-121

examples, II-122, II-126, II-132, II-145, II-146, II-148, II-151, II-155, II-160, II-163, II-166, II-170

factorial, II-24

homogeneity tests, II-113

hypothesis tests, II-23, II-113, II-115, II-116

interactions, II-25

normality tests, II-112

pairwise comparisons, II-117

power analysis, IV-19, IV-26, IV-55, IV-57, IV-77, IV-80

Quick Graphs, II-121

repeated measures, II-31, II-110

resampling, II-108

residuals, II-110

sums of squares, II-113

two-way ANOVA, IV-26, IV-57, IV-80

unbalanced designs, II-29

unequal variances, II-26

usage, II-121

within-subject differences, II-32

Anderberg dichotomy coefficients, I-164, I-173

Anderberg's binary similarity coefficient, I-164

Anderson-Darling test, I-303

Andrews procedure, III-279

angle tolerance, IV-388

anisotropy, IV-392, IV-405

geometric, IV-392

zonal, IV-393

A-optimality, I-364

ARIMA models, IV-514, IV-523, IV-540

algorithms, IV-578

arithmetic mean, I-299, I-308

ARMA models, IV-519

asymptotically distribution-free estimates, III-412

autocorrelation plots, II-11, IV-516, IV-520

Automatic Interaction Detection(AID), I-45, I-47

autoregressive models, IV-516

average run length curves, IV-134

chart types, IV-137

continuous distributions, IV-139

discrete distributions, IV-140

overview, IV-134

probability limits, IV-137

axial designs, I-360

B

backward elimination, II-15
 bandwidth, IV-350, IV-355, IV-388
 optimal values, IV-356
 relationship with kernel function, IV-357
 basic statistics
 Anderson-Darling test, I-303, I-309
 columns, I-307
 commands, I-322
 Cronbach's alpha, I-321
 examples, I-324, I-326, I-327, I-328, I-333, I-338, I-340, I-341, I-342
 geometric mean, I-300, I-308
 harmonic mean, I-300, I-308
 multivariate normality assessment, I-303
 N-&P-tiles, I-309
 overview, I-297
 Quick Graphs, I-323
 resampling, I-298
 rows, I-316
 Shapiro-Wilk test, I-302, I-309
 stem-and-leaf for columns, I-314
 stem-and-leaf for rows, I-320
 test for normality, I-302
 trimmed mean, I-299, I-308
 usage, I-323
 bayesian regression, II-50
 credibility intervals, II-50
 gamma prior, II-52
 normal prior, II-52
 best linear unbiased estimates (BLUE), II-344, II-386
 best linear unbiased predictors (BLUP), II-344, II-386
 beta level, IV-22
 between-group differences
 in analysis of variance, II-32
 bias, II-15
 binary logit, III-2
 compared to multinomial logit, III-5
 binary trees, I-43
 biplots, IV-6, IV-8

bisquare procedure, III-279
 biweight kernel, IV-365
 Bonferroni inequality, I-47
 Bonferroni test, I-175, II-27, II-118, II-196, II-307, II-394
 bootstrap, I-19, I-21
 box plot, I-305
 Box-and-Whisker plots, IV-112
 Box-Behnken designs, I-357, I-380
 Box-Cox power transformation, IV-157
 Box-Hunter designs, I-353, I-373
 Bray-Curtis measure, I-162, I-172
 broad inference space, II-280

C

c charts, IV-131
 C matrix, II-193
 candidate sets
 for optimal designs, I-363
 canonical correlation analysis
 data format, IV-304
 examples, IV-305, IV-308, IV-312
 interactions, IV-304
 model, IV-299
 nominal scales, IV-304
 overview, IV-291
 partialled variables, IV-300
 Quick Graphs, IV-305
 resampling, IV-291
 rotation, IV-303
 usage, IV-304
 canonical rotation, IV-7
 categorical data, III-321
 categorical predictors, I-45
 Cauchy kernel, IV-365
 CCF plots, IV-531
 central composite designs, I-356, I-384
 centroid designs, I-359
 CHAID, I-46, I-47
 chi-square tests for independence, I-229, I-233, I-242
 circle model

- in perceptual mapping, IV-5
- city-block distance, I-172, III-191
- classical analysis, IV-488
- classification and regression trees, I-41
- classification functions, I-396
- classification trees
 - algorithms, I-62
 - basic tree model, I-42
 - commands, I-54
 - compared to discriminant analysis, I-46, I-49, I-46
 - data format, I-54
 - displays, I-51
 - examples, I-55, I-57, I-59
 - loss functions, I-51
 - missing data, I-62
 - mobiles, I-41
 - model, I-51
 - overview, I-41
 - pruning, I-47
 - Quick Graphs, I-54
 - resampling, I-41
 - saving files, I-54
 - stopping criteria, I-47, I-53
 - usage, I-54
- cluster analysis
 - additive trees, I-91
 - algorithms, I-122
 - clustering, I-65
 - commands, I-93
 - data types, I-95
 - distances, I-84
 - examples, I-96, I-105, I-108, I-109, I-111, I-112, I-115, I-116, I-118, I-120
 - exclusive clusters, I-66
 - hierarchical clustering, I-82
 - k-means clustering, I-78
 - k-medians clustering, I-79
 - missing values, I-122
 - overlapping clusters, I-66
 - overview, I-65
 - Quick Graphs, I-95
 - resampling, I-66
 - saving files, I-95
 - usage, I-95
- clustered data, II-421
- clustering
 - hierarchical clustering, I-68
 - k-clustering, I-78
 - validity, I-87
- Cochran's test of linear trend, I-234
- coefficient of alienation, III-190, III-212
- coefficient of determination
 - see multiple correlation
- coefficient of variation, I-307
- Cohen's kappa, I-226, I-234
- communalities, I-458
- compound symmetry, II-32
- conditional logistic regression, III-5
- confidence curves, III-273
- confidence intervals, I-11, I-307
 - path analysis, III-455
- conjoint analysis
 - additive tables, I-126
 - algorithms, I-152
 - commands, I-135
 - compared to logistic regression, I-132
 - data format, I-135
 - examples, I-136, I-140, I-143, I-147
 - missing data, I-153
 - model, I-133
 - multiplicative tables, I-128
 - overview, I-125
 - Quick Graphs, I-135
 - resampling, I-125
 - saving files, I-135
 - usage, I-135
- constraints
 - in mixture designs, I-360
- contingency coefficient, I-227
- contour plot, IV-243
- contour plots, IV-401
- contrast coefficients, II-31
- contrasts
 - in analysis of variance, II-28
- control charts

- aggregated data, IV-120
- average run length curves, IV-136
- control limits, IV-121
- discrete control limits, IV-121
- operating characteristic curves, IV-135
- raw data, IV-120
- regression charts, IV-152
- sigma limits, IV-122
- convergence, III-98
- convex hulls, IV-398
- Cook's distance, II-12
- Cook-Weisberg graphical confidence curves, III-273
- coordinate exchange method, I-363, I-386
- correlations, I-67, I-157
 - algorithms, I-199
 - binary data, I-173
 - canonical, IV-291
 - commands, I-177
 - continuous data, I-171
 - data format, I-178
 - dissimilarity measures, I-172
 - distance measures, I-172
 - examples, I-179, I-182, I-185, I-186, I-188, I-192, I-195, I-196, I-198
 - missing values, I-170, I-199, III-135
 - options, I-174
 - overview, I-157
 - power analysis, IV-19, IV-25, IV-42, IV-44
 - Quick Graphs, I-178
 - rank-order data, I-172
 - resampling, I-158
 - saving files, I-179
 - set, IV-291
 - usage, I-178
- correlograms, IV-403
- correspondence analysis, IV-2, IV-6
 - algorithms, I-218
 - commands, I-206
 - data format, I-206
 - examples, I-207, I-214
 - missing data, I-218
 - model, I-204
 - overview, I-201
 - Quick Graphs, I-206
 - resampling, I-201
 - simple correspondence analysis, I-204
 - usage, I-206
- covariance matrix, I-171, III-135
- covariance paths
 - path analysis, III-401
- covariograms, IV-387
- Cox-Snell residual plot, IV-434
- Cramer's V, I-227
- critical level, I-13
- Cronbach's alpha, IV-488, IV-489
 - see basic statistics, I-321
- crossover designs, II-175
- crosstabulation
 - commands, I-244
 - data format, I-246
 - examples, I-248, I-250, I-253, I-256, I-257, I-258, I-261, I-263, I-269, I-271, I-273, I-275, I-277, I-279, I-293
 - multiway, I-237
 - one-way, I-220, I-222, I-228
 - overview, I-219
 - Quick Graphs, I-247
 - resampling, I-219
 - standardizing tables, I-221
 - two-way, I-220, I-223, I-231
 - usage, I-246
- cross-validation, I-48, I-396, II-16, III-360
- cumulative sum charts
 - see cusum charts, IV-142
- D
 - D matrix, II-194, II-288, II-309, II-355, II-397
 - D SUB-A (d_a), IV-321
 - dates, IV-430
 - dendrograms, I-65, I-107
 - dependence paths
 - path analysis, III-399
 - descriptive statistics, I-1
 - see basic statistics, I-307

- design of experiments, I-132, I-368, I-369
 - axial designs, I-360
 - Box-Behnken designs, I-357
 - central composite designs, I-356
 - centroid designs, I-359
 - commands, I-370
 - examples, I-371, I-372, I-373, I-375, I-377, I-379, I-380, I-381, I-382, I-384, I-386
 - factorial designs, I-349, I-350
 - lattice designs, I-359
 - mixture designs, I-350, I-357
 - optimal designs, I-350, I-362
 - overview, I-345
 - Quick Graphs, I-371
 - response surface designs, I-350, I-354
 - screening designs, I-360
 - usage, I-370
- determinant criterion
 - see D-optimality
- Dice's binary similarity coefficient, I-164
- dichotomy coefficients, I-164
 - Anderberg, I-173
 - Jaccard, I-173
 - positive matching, I-173
 - simple matching, I-173
 - Tanimoto, I-173
- difficulty, IV-507
- discrete choice model, III-7
 - compared to polytomous logit, III-8
- discrete gaussian convolution, IV-361
- discriminant analysis
 - classical discriminant analysis, I-400
 - commands, I-407
 - data format, I-408
 - estimation, I-401
 - examples, I-409, I-413, I-420, I-427, I-435, I-438, I-444, I-449
 - linear discriminant function, I-397
 - model, I-400
 - multiple groups, I-399
 - options, I-401
 - overview, I-391
 - prior probabilities, I-398
 - Quick Graphs, I-408
 - resampling, I-391
 - robust discriminant analysis, I-399
 - statistics, I-404
 - stepwise estimation, I-401
 - usage, I-408
- discrimination parameter, IV-507
- dissimilarities
 - direct, III-187
 - indirect, III-187
- distance measures, I-67, I-157
- distances
 - nearest-neighbor, IV-396
- distance-weighted least squares (DWLS) smoother, IV-361
- distributions
 - Benford's law, I-499, III-332, IV-86, IV-221
 - beta, I-500, III-333, III-335, IV-88, IV-222
 - binomial, I-499, III-332, IV-86, IV-221
 - Cauchy, I-500, III-333, III-335, IV-88, IV-222
 - chi-square, I-500, III-333, III-335, IV-88, IV-222
 - discrete uniform, I-499, III-332, IV-86, IV-221
 - double exponential, I-501, III-335, IV-88, IV-222
 - Erlang, I-501, III-335, IV-88, IV-222
 - exponential, I-501, III-333, III-336, IV-88, IV-222
 - F, III-333, III-336, IV-88, IV-222
 - gamma, I-501, III-333, III-336, IV-89, IV-222
 - generalized lambda, IV-222
 - geometric, I-499, III-332, IV-86, IV-221
 - Gompertz, I-501, III-333, III-336, IV-89, IV-222
 - Gumbel, I-501, III-333, III-336, IV-89, IV-222
 - hypergeometric, I-499, III-332, IV-86, IV-221
 - inverse Gaussian, I-501, III-333, III-336, IV-89, IV-222
 - logarithmic series, I-499, III-332, IV-87, IV-221
 - logistic, I-501, III-333, III-336, IV-89, IV-222

- logit normal, I-501, III-333, III-336, IV-89, IV-222
- loglogistic, I-501, III-333, III-336, IV-89, IV-222
- lognormal, I-501, III-333, III-336, IV-89, IV-222
- negative binomial, I-499, III-333, IV-87, IV-221
- non-central chi-square, III-333, III-336, IV-89, IV-222
- non-central F, III-333, III-336, IV-89, IV-222
- non-central t, III-333, III-336, IV-89, IV-222
- normal, I-501, III-333, III-336, IV-89, IV-222
- Pareto, I-501, III-333, III-336, IV-89, IV-222
- Poisson, I-499, III-333, IV-87, IV-221
- Rayleigh, I-501, III-333, III-336, IV-89, IV-223
- smallest extreme value, I-501, III-333, III-336, IV-89, IV-223
- studentized maximum modulus, III-333, III-336, IV-89
- Studentized range, III-336
- studentized range, III-333, IV-89, IV-223
- t, III-333, III-336, IV-89, IV-223
- triangular, I-501, III-334, III-336, IV-89, IV-223
- uniform, I-501, III-334, III-336, IV-89, IV-223
- Weibull, I-501, III-334, III-336, IV-89, IV-223
- zipf, I-499, III-333, IV-87, IV-221
- dit plots, I-14
- D-optimality, I-364
- dot histogram plots, I-14
- Double, III-333
- D-Prime (d'), IV-320
- dummy codes, II-180
- Duncan test, II-27, II-119, II-197
- Dunnett test, II-27, II-119, II-197
- Dunnett's T3 test, II-27, II-119, II-197
- Dunn-Sidak test, I-175
- E
- ECVI, III-458
- edge effects, IV-398
- effect size
 - in power analysis, IV-22, IV-23
- effects coding, II-20, II-180
- efficiency, I-362
- eigenvalues, I-405
- ellipse model
 - in perceptual mapping, IV-6
- EM algorithm, I-492
- EM estimation, III-130
 - for correlations, I-175, III-135
 - for covariance, III-135
 - for SSCP matrix, III-135
- endogenous variables
 - path analysis, III-400
- Epanechnikov kernel, IV-364
- equamax rotation, I-460, I-464
- Erlang, III-333
- Estimation, III-135
- Euclidean distances, III-188
- exogenous variables
 - path analysis, III-400
- expected cross-validation index, III-458
- Exponential, III-336
- exponential distribution, IV-432
- exponential model, IV-390, IV-404
- exponential smoothing, IV-524
- exponentially weighted moving average charts, IV-146
 - control limits, IV-147
- external unfolding, IV-4
- F
- F, III-333
- F and R matrices, II-308, II-354, II-396
- F distribution
- F matrix, II-287
- factor analysis, I-457, IV-2
 - algorithms, I-492
 - commands, I-468

- compared to principal components analysis, I-460
 - convergence, I-463
 - correlations vs covariances, I-457
 - eigenvalues, I-463
 - eigenvectors, I-467
 - examples, I-469, I-473, I-476, I-478, I-482, I-485
 - iterated principal axis, I-463
 - loadings, I-467
 - maximum likelihood, I-463
 - missing values, I-492
 - number of factors, I-463
 - overview, I-453
 - principal components, I-463
 - Quick Graphs, I-468
 - resampling, I-453
 - residuals, I-465
 - rotation, I-459, I-464
 - save, I-466
 - scores, I-466
 - usage, I-468
 - factor loadings, IV-488
 - factorial analysis of variance, II-24
 - factorial designs, I-349, I-350
 - analysis of, I-353
 - examples, I-371
 - fractional factorials, I-352
 - full factorial designs, I-352
 - F-distribution
 - non-centrality parameter, IV-60
 - Fedorov method, I-363
 - Fieller bounds, III-48
 - filters, IV-527
 - Fisher's exact test, I-226, I-233
 - Fisher's linear discriminant function, IV-2
 - Fisher's LSD, II-197
 - Fisher's LSD test, II-27, II-118, II-307, II-395
 - fitting distributions
 - commands, I-501
 - examples, I-504, I-505, I-507, I-508, I-510, I-511, I-513
 - goodness-of-fit tests, I-496
 - maximum likelihood method, I-497
 - method of moments, I-497
 - method of quantiles or order statistic, I-497
 - overview, I-495
 - Quick Graphs, I-503
 - Shapiro-Wilk's test for normality, I-497
 - usage, I-503
 - fixed effects, II-279
 - fixed variance
 - path analysis, III-402
 - fixed-bandwidth method
 - compared to KNN method, IV-357
 - for smoothing, IV-355, IV-357, IV-364
 - Fletcher-Powell minimization, IV-507
 - forward selection, II-15
 - Fourier analysis, IV-526, IV-545
 - fractional factorial designs
 - Box-Hunter designs, I-353
 - examples, I-372, I-373, I-375, I-377, I-379
 - homogeneous fractional designs, I-353
 - Latin square designs, I-353
 - mixed-level fractional designs, I-353
 - Plackett-Burman designs, I-353
 - Taguchi designs, I-353
 - Freeman-Tukey deviates, III-93, III-102
 - frequencies, I-23, I-54, I-135, I-179, I-206, I-246, I-248, I-323, I-408, I-468, I-469, I-503, I-544, II-54, II-121, II-122, II-202, II-310, II-357, II-399, II-441, III-23, III-103, III-104, III-137, III-194, III-217, III-283, III-339, III-364, III-385, III-413, IV-9, IV-62, IV-63, IV-103, IV-162, IV-244, IV-280, IV-305, IV-325, IV-328, IV-366, IV-410, IV-449, IV-495, IV-498, IV-547, IV-587
 - frequency tables, III-93, III-102
 - see crosstabulation
 - Friedman test, III-328
- ## G
- Gabriel test, II-27, II-119, II-197
 - Games-Howell test, II-27, II-119, II-197
 - Gaussian kernel, IV-364, IV-365
 - Gaussian model, IV-390, IV-404

Gauss-Newton method, III-269, III-272
 general linear models, II-175
 algorithms, II-249
 categorical variables, II-179
 commands, II-200
 contrasts, II-189, II-191
 data format, II-201
 examples, II-203, II-211, II-212, II-213, II-215, II-217, II-220, II-222, II-224, II-234, II-237, II-238, II-242, II-246, II-247, II-248
 hypothesis options, II-188
 hypothesis tests, II-186
 mixture model, II-184
 model estimation, II-177
 overview, II-175
 pairwise comparisons, II-195
 post hoc tests, II-199
 Quick Graphs, II-202
 resampling, II-176
 stepwise regression, II-183
 usage, II-201
 generalized least squares, III-412, IV-584
 generalized variance, IV-294
 geometric mean, I-300, I-308
 geostatistical models, IV-386, IV-387
 getween-groups testing, III-239
 Gini index, I-48, I-51
 GLM
 see general linear models, II-175
 global criterion
 see G-optimality
 GMA chart, IV-146
 Goodman-Kruskal gamma, I-227, I-234
 Goodman-Kruskal lambda, I-234
 goodness-of-fit tests, I-496
 G-optimality, I-364
 Gower2 binary similarity coefficient, I-164
 Graeco-Latin square designs, I-353
 Greenhouse-Geisser statistic, II-33
 Guttman μ_2 monotonicity coefficients, I-162
 Guttman's coefficient of alienation, III-190
 Guttman's loss function, III-212

Guttman-Rulon coefficient, IV-489

H

Hadi outlier detection, I-168
 Hamman's binary similarity coefficient, I-164
 Hampel procedure, III-279
 Hanning weights, IV-512
 harmonic mean, I-300, I-308
 hazard function
 heterogeneity, IV-435
 Henderson's mixed model equations, II-279, II-293
 Henze-Zirkler test, I-303
 heteroskedasticity, IV-583
 heteroskedasticity-consistent standard errors, IV-583
 hierarchical clustering, I-68, I-82
 distances, I-84
 validity index, I-75
 hierarchical linear mixed models
 categorical variables, II-389
 commands, II-398
 examples, II-399, II-402, II-406, II-408, II-412, II-414, II-417
 hypothesis testing, II-394
 model estimation, II-387
 options, II-392
 overview, II-385
 Quick Graphs, II-398
 random effects, II-390
 usage, II-398
 hierarchical linear models
 see mixed regression
 hinge, I-301
 Hochberg's GT2 test, II-27, II-119, II-197, II-307, II-395
 hole model, IV-391, IV-405
 Holt's method, IV-524
 homogeneity tests, II-113
 Levene's test, II-113
 Hotelling's T squared charts, IV-153
 Hotelling-Lawley trace, III-226
 Huber procedure, III-279

Huynh-Feldt statistic, II-33
 hyper-Graeco-Latin square designs, I-353
 hypothesis

alternative, I-13

null, I-13

testing, I-12, II-7

hypothesis testing

Bartlett's test, I-521

commands, I-541

confidence intervals, I-520, I-521, I-522

data format, I-543

examples, I-544, I-545, I-547, I-548, I-549, I-551, I-552, I-556, I-557, I-560, I-562, I-564

Levene's tests, I-521

multiple tests, I-522

overview, I-519

Quick Graphs, I-544

resampling, I-519

test for means, I-520

tests for correlation, I-522

tests for mean, I-520

tests for proportion, I-520, I-538

tests for variance, I-521

usage, I-543

I

ID3, I-47

I-MR chart

see X-MR chart, IV-150

incomplete block designs, II-175

independence, I-223

in loglinear models, III-94

individual cases charts

See X charts, IV-129

INDSCAL model, III-185

inertia, I-202

inferential statistics, I-7, IV-20

instrumental variables, IV-582

intermediate inference space, II-280

internal-consistency, IV-489

interquartile range, I-301

interval censored data, IV-428

inverse-distance smoother, IV-360

isotropic, IV-387

item-response analysis

see test item analysis

item-test correlations, IV-488

J

Jaccard dichotomy coefficients, I-164, I-173

jackknife, I-18, I-22

jackknifed classification matrix, I-396

K

k nearest-neighbors method

compared to fixed-bandwidth method, IV-357

for smoothing, IV-356, IV-362

k-clustering, I-78

k-means, I-78

k-medians, I-79

Kendall's Tau b, I-172

Kendall's tau-b coefficient, I-227

kernel functions, IV-350, IV-352

biweight, IV-364

Cauchy, IV-364

Epanechnikov, IV-364

Gaussian, IV-364

plotting, IV-354

relationship with bandwidth, IV-357

tricube, IV-364

triweight, IV-362, IV-364

k-exchange method, I-363

Kolmogorov-Smirnov test, III-319

KR20, IV-489

kriging, IV-405

ordinary, IV-394, IV-405, IV-407

simple, IV-393, IV-407

trend components, IV-394

universal, IV-394, IV-407

Kruskal's loss function, III-211

Kruskal's STRESS, III-190

Kruskal-Wallis test, III-319

K-S test, III-319

- Kulczynski's binary similarity coefficient, I-164
 Kulczynski's binary similarity coefficient, I-173
 kurtosis, I-307
- L**
- latent trait model, IV-488, IV-490
 Latin square designs, I-353, I-375
 lattice, III-382
 lattice designs, I-359
 least absolute deviations, III-268
 least absolute deviations regression, IV-260
 least median of squares regression, IV-261
 search method, IV-269
 least trimmed squares regression, IV-261
 Levene test, II-25
 leverage, II-12
 likelihood ratio chi-square, I-233, III-96, III-101
 compared to Pearson chi-square, III-96
 likelihood-ratio chi-square, I-226
 Lilliefors test, III-334, III-355
 linear contrasts, II-28
 linear discriminant model, I-392
 linear mixed models
 categorical variables, II-347
 commands, II-356
 examples, II-357, II-362, II-366, II-369, II-372, II-379, II-382
 hypothesis testing, II-352
 model estimation, II-345
 options, II-350
 overview
 Quick Graphs, II-356
 random effects, II-348
 usage, II-356
 linear models
 general linear models, II-175
 hierarchical, II-421
 linear discriminant model, I-392
 linear regression, II-39, II-299, II-385
 linear regression, I-11, II-7, II-39
 AIC and Schwarz's BIC, II-39
 Anderson-Darling test, II-45
 bayesian, II-50
 commands, II-53
 data format, II-54
 examples, II-55, II-60, II-63, II-67, II-71, II-75, II-81, II-83, II-85, II-86, II-87, II-89, II-95, II-97, II-99
 Kolmogorov-Smirnov test, II-45
 model, II-41
 normality tests, II-45
 overview, II-39
 prediction intervals, II-40, II-46
 Quick Graphs, II-54
 resampling, II-40, II-47
 residuals, II-9, II-41
 ridge, II-48
 Shapiro-Wilk test, II-45
 stepwise, II-15
 tolerance, II-43
 usage, II-54
 using correlation matrix as input, II-18, II-89
 using covariance matrix as input, II-18, II-89
 using SSCP matrix as input, II-18, II-89
 variance inflation factor, II-70
 listwise deletion, I-492, III-125
 Little's MCAR test, III-123, III-133
 loadings, I-456, I-457
 LOESS smoothing, IV-361, IV-363, IV-367, IV-368, IV-370, IV-380
 logistic item-response analysis, IV-506
 one-parameter model, IV-490
 two-parameter model, IV-490
 logistic regression
 AIC and Schwarz's BIC, III-1
 algorithms, III-85
 categorical predictors, III-11
 classification table, III-17
 compared to conjoint analysis, I-132
 conditional variables, III-10
 confidence intervals, III-48
 data format, III-22
 deciles of risk, III-17
 discrete choice, III-13
 dummy coding, III-11, III-12

- effect coding, III-11, III-12
- estimation, III-15
- examples, III-24, III-27, III-33, III-39, III-45, III-50, III-60, III-69, III-70, III-77, III-81
- missing data, III-86
- model, III-10
- options, III-14
- overview, III-1
- post hoc tests, III-20
- prediction table, III-16
- quantiles, III-18, III-49
- Quick Graphs, III-23
- regression diagnostics, III-87
- robust standard errors, III-16
- ROC curve, III-1
- simulation, III-19
- usage, III-22
- weights, III-23
- logit
 - binary logit, III-2
 - conditional logit, III-5
 - discrete choice logit, III-7
 - multinomial logit, III-5
 - stepwise logit, III-9
- loglinear modeling
 - commands, III-103
 - compared to analysis of variance, III-95
 - compared to Crosstabs, III-102
 - convergence, III-96
 - data format, III-103
 - examples, III-105, III-114, III-117, III-121
 - frequency tables, III-102
 - model, III-96
 - overview, III-93
 - parameters, III-100
 - Quick Graphs, III-104
 - saturated models, III-95
 - statistics, III-100
 - structural zeros, III-98
 - usage, III-103
- log-logistic distribution, IV-432
- lognormal distribution, IV-432
- longitudinal data, II-421
- loss function, III-265
 - multidimensional scaling, III-210
- loss functions, I-48
- LOWESS smoothing, IV-513
- low-pass filter, IV-527
- LSD test, II-197
- M
 - madograms, IV-403
 - Mahalanobis distances, I-392
 - Mann-Whitney, III-342
 - Mantel-Haenszel test, I-238
 - Mardia skewness and kurtosis, I-298, I-303
 - Marquardt method, III-275
 - Marron & Nolan canonical kernel width, IV-357, IV-364
 - mass, I-202
 - matrix displays, I-70
 - maximum likelihood estimates, II-385, III-266
 - maximum likelihood factor analysis, I-461
 - Maximum Wishart likelihood, III-411
 - McFadden's conditional logit model, III-7
 - McNemar's test, I-226, I-234
 - MDPREF, IV-6, IV-8
 - MDS
 - see multidimensional scaling, III-185
 - mean, I-3, I-307
 - mean smoothing, IV-358, IV-365
 - means coding, II-21
 - median, I-4, I-299, I-307
 - median smoothing, IV-358
 - meta-analysis, II-19
 - midrange, I-301
 - minimum spanning trees, IV-396
 - Minkowski metric, III-191
 - MIS function, III-142
 - Missing At Random(MAR), III-131
 - Missing Completely At Random(MCAR), III-131
 - missing value analysis
 - casewise pattern table, III-142
 - data format, III-137

- EM algorithm, III-130, III-134, III-135, III-154, III-168, III-176
- examples, III-137, III-142, III-154, III-168, III-176
- listwise deletion, III-125, III-154, III-168
- MISSING command, III-136
- missing value patterns, III-137
- model, III-134
- outliers, III-135
- overview, III-123
- pairwise deletion, III-125, III-154, III-168
- pattern variables, III-124, III-176
- Quick Graphs, III-137
- randomness, III-131
- regression imputation, III-127, III-134, III-154, III-176
- resampling, III-123
- saving estimates, III-134, III-137
- unconditional mean imputation, III-126
- usage, III-137
- mixed models, II-251
 - AIC and Schwarz's BIC, II-292
 - ANOVA Method, II-281
 - compound symmetry structure, II-270
 - covariance structures, II-269
 - diagonal structure, II-271
 - estimation methods, II-281
 - hypothesis testing, II-286
 - MIVQUE(0) method, II-283
 - ML method, II-284
 - pairwise comparison, II-290
 - post hoc tests, II-290
 - REML method, II-285
 - setup, II-267
 - unstructured (general symmetric structure), II-272
 - variance components structure, II-270
- mixed regression
 - algorithms, II-484
 - commands, II-441
 - data format, II-441
 - examples, II-442, II-449, II-457, II-473
 - overview, II-421
 - Quick Graphs, II-441
 - usage, II-441
- mixture designs, I-350, I-357
 - analysis of, I-361
 - axial designs, I-360
 - centroid designs, I-359
 - constraints, I-360
 - examples, I-381, I-382
 - lattice designs, I-359
 - Scheffé model, I-361
 - screening designs, I-360
 - simplex, I-359
- models, I-10, II-301
 - estimation, I-10
- moving average, IV-355, IV-511, IV-517
- moving average chart, IV-144
- moving-averages smoother, IV-360
- M-regression, IV-261
- multidimensional scaling, III-185, IV-2
 - algorithms, III-211
 - assumptions, III-186
 - commands, III-194
 - configuration, III-189, III-193
 - confirmatory, III-193
 - convergence, III-192
 - data format, III-194
 - dissimilarities, III-187
 - distance metric, III-189
 - examples, III-195, III-198, III-200, III-203, III-208
 - Guttman method, III-212
 - individual differences, III-185
 - Kruskal method, III-211
 - log function, III-191
 - loss function, III-190
 - metric, III-189
 - missing values, III-212
 - nonmetric, III-189
 - overview, III-185
 - power function, III-191
 - Quick Graphs, III-194
 - residuals, III-192
 - R-metric, III-191

- Shepard diagrams, III-189, III-194
- usage, III-194
- multilevel models
 - see mixed regression
- multinomial logit, III-5
 - compared to binary logit, III-5
- multinormal tests, III-215
 - examples, III-218, III-219
 - Henze-Zirkler test, III-215
 - Mardia skewness and kurtosis, III-215
 - overview, III-215
 - Quick Graphs, III-217
 - usage, III-217
 - using commands, III-217
- multiple comparison tests
 - see pairwise comparisons, II-117, II-195
- multiple correlation, II-8
- multiple correspondence analysis, I-203
- multiple regression, II-12
- multiple tests
 - Bonferroni adjustment, I-522
 - Dunn-Sidak adjustment, I-522
- multivariate analysis of variance, III-223
 - between-groups testing, III-239
 - categorical variables, III-229
 - commands, III-244
 - data format, III-244
 - examples, III-246, III-248, III-253, III-255, III-257, III-258
 - Hotelling-Lawley trace, III-226
 - hypothesis test, III-232
 - overview, III-223
 - Pillai trace, III-225
 - post hoc test, III-242
 - Quick Graphs, III-245
 - repeated measures, III-230
 - Roy's Greatest root, III-226
 - usage, III-244
 - Wilks' lambda, III-225
 - within-group testing, III-241
- multivariate normality assessment
 - Henze-Zirkler test, I-303
 - Mardia's skewness, I-303
- mutually exclusive, I-222
- N
- N- & P-tiles, I-309
 - methods, I-311
 - transformation, I-309
- Nadaraya-Watson smoother, IV-360
- narrow inference space, II-280
- Nelson-Aalen cumulative hazard estimator, IV-438
- nesting, II-175
- Newton-Raphson method, III-93
- NIPALS (Nonlinear Iterative Partial Least Squares)
 - see partial least squares regression, III-377
- nodes, I-43
- nominal data, III-321
- non-central F-distribution, IV-34, IV-60
- non-centrality parameters, IV-34
- nonlinear models, III-261
 - algorithms, III-316
 - commands, III-283
 - computation, III-274, III-316
 - convergence, III-274, III-275
 - data format, III-283
 - estimation, III-269
 - examples, III-284, III-287, III-290, III-293, III-296, III-298, III-299, III-301, III-306, III-311, III-313, III-315
 - functions of parameters, III-277
 - loss functions, III-265, III-270, III-280, III-281
 - missing data, III-316
 - model, III-270
 - parameter bounds, III-274
 - problems, III-269
 - Quick Graphs, III-283
 - recalculation of parameters, III-276
 - resampling, III-261
 - robust estimation, III-278
 - starting values, III-274
 - usage, III-283
- nonmetric unfolding model, III-185
- nonparametric statistics, III-325

nonparametric tests

- algorithms, III-355
 - Anderson-Darling test, III-334
 - commands, III-325, III-331, III-338
 - data format, III-339
 - examples, III-340, III-342, III-343, III-345, III-346, III-347, III-348, III-349, III-350, III-353, III-354
 - Friedman test, III-328
 - independent samples test, III-322, III-323
 - Kolmogorov-Smirnov test, III-323, III-331
 - Kruskal-Wallis test, III-322
 - Mann-Whitney test, III-322
 - overview, III-319
 - Quade test, III-329
 - Quick Graphs, III-339
 - related variables tests, III-325, III-326, III-328
 - resampling, III-319
 - sign test, III-325, III-326
 - usage, III-339
 - Wald-Wolfowitz runs test, III-337
 - Wilcoxon Signed-Rank test, III-326
- normal distribution, I-301
- normality tests, II-45, II-112
- Anderson-Darling, II-113
 - Anderson-Darling test, II-45
 - Kolmogorov-Smirnov test, II-45, II-112
 - Shapiro-Wilk, II-112
 - Shapiro-Wilk test, II-45
- np charts, IV-129
- NPAR, IV-320
- null hypothesis, I-12, IV-20

O

- oblimin rotation, I-460, I-464
- observational studies, I-347
- OC curves, IV-134
- Occam's razor, I-130
- Ochiai's binary similarity coefficient, I-164
- odds ratio, I-233
- omni-directional variograms, IV-388
- operating characteristic curves

- chart type, IV-136
 - continuous distributions, IV-139
 - discrete distributions, IV-140
 - overview, IV-134
 - probability limits, IV-136
 - sample size, IV-138
 - scaling, IV-138
- optimal designs, I-350, I-362
- analysis of, I-364
 - A-optimality, I-364
 - candidate sets, I-363
 - coordinate exchange method, I-363, I-386
 - D-optimality, I-364
 - efficiency criteria, I-364
 - Fedorov method, I-363
 - G-optimality, I-364
 - k-exchange method, I-363
 - model, I-365
 - optimality criteria, I-364
- optimality, I-362
- ORDER, IV-431
- ordinal data, III-320
- Ordinary least squares, III-412
- orthomax rotation, I-460, I-464
- Output, IV-99

P

- p charts, IV-130
- PACF plots, IV-530
- pairwise comparisons, II-26, II-107, II-117
 - Bonferroni test, II-118, II-196
 - Duncan test, II-119, II-197
 - Dunnett test, II-119, II-197
 - Dunnett's T3 test, II-119, II-197
 - Fisher's LSD, II-197
 - Fisher's LSD test, II-118
 - Gabriel test, II-119, II-197
 - Games - Howell test, II-197
 - Games-Howell test, II-119
 - Hochberg's GT2 test, II-119
 - Hochberg's test GT2, II-197
 - R-E-G-W Q test, II-197

- R-E-G-W-Q test, II-119
- Scheffé test, II-27, II-118, II-197
- Sidak test, II-118, II-197
- Student-Newman-Keuls test, II-119, II-197
- Tamhane's T2 test, II-119, II-197
- Tukey test, II-118, II-196
- Tukey's b test, II-119, II-197
- pairwise deletion, I-492, III-125
- parameters, I-10
- parametric modeling, IV-432
- Pareto charts, IV-111
- partial autocorrelation plots, IV-519, IV-520
- partial least squares regression
 - algorithms, III-377
 - cross-validation, III-363
 - examples, III-365, III-368, III-371, III-375
 - latent factors, III-357, III-359
 - leave-one-out, III-360, III-363
 - NIPALS, III-362
 - PRESS statistic, III-360
 - Quick Graphs, III-364
 - random exclusion, III-360, III-364
 - SIMPLS, III-362
 - test set, III-360
 - training set, III-360
 - usage, III-364
 - using commands, III-364
- partialing
 - in set correlation, IV-295
- partially ordered scalogram analysis with coordinates
 - algorithms, III-395
 - commands, III-385
 - Convergence, III-384
 - convergence, III-384
 - data format, III-385
 - displays, III-383
 - examples, III-386, III-388, III-390
 - missing data, III-395
 - model, III-384
 - overview, III-381
 - Quick Graphs, III-385
 - resampling, III-381
 - usage, III-385
- path analysis
 - algorithms, III-454
 - confidence intervals, III-455
 - covariance paths, III-401
 - covariance relationship, III-409
 - data format, III-413
 - dependence paths, III-399
 - dependence relationship, III-407
 - endogenous variables, III-400
 - estimate, III-411
 - examples, III-414, III-419, III-434, III-442
 - exogenous variables, III-400
 - fixed variance, III-402
 - free parameters, III-418
 - latent variables, III-404
 - manifest variables, III-410
 - measures of fit, III-455
 - method of estimation, III-411
 - model, III-452
 - model statement, III-407
 - options, III-411
 - overview, III-397
 - path diagrams, III-397
 - Quick Graphs, III-413
 - starting values, III-412
 - usage, III-413
 - variance paths, III-401
- Pearson chi-square, I-223, I-228, I-233, III-94, III-101
 - compared to likelihood ratio chi-square, III-96
- Pearson correlation, I-160, I-171
- perceptual mapping
 - algorithms, IV-16
 - commands, IV-9
 - data format, IV-9
 - examples, IV-9, IV-11, IV-12, IV-14
 - methods, IV-8
 - missing data, IV-16
 - model, IV-7
 - overview, IV-1
 - PREFMAP, IV-1
 - Quick Graphs, IV-9

- usage, IV-9
- periodograms, IV-527
- permutation tests, I-222
- phi coefficient, I-48, I-51, I-52, I-227
- Pillai trace, III-225
- Plackett-Burman designs, I-353, I-379
- point processes, IV-386, IV-395
- polynomial contrasts, II-28, II-31, II-192
- polynomial smoothing, IV-358, IV-365
- populations, I-7
- POSET, III-381
- positive matching dichotomy coefficients, I-164, I-173
- Post hoc Test for Repeated measures, III-242
- power, IV-22
- power analysis
 - analysis of variance, IV-19
 - commands, IV-62
 - correlation coefficients, IV-25, IV-42, IV-44
 - correlations, IV-19
 - data format, IV-62
 - examples, IV-63, IV-67, IV-72, IV-77, IV-80
 - generic, IV-34, IV-60, IV-77
 - one-sample t-test, IV-26
 - one-sample z-test, IV-46
 - one-way ANOVA, IV-26, IV-55, IV-77
 - overview, IV-19
 - paired t-test, IV-26, IV-51, IV-67
 - power curves, IV-62
 - proportions, IV-19, IV-25, IV-39, IV-40, IV-63
 - Quick Graphs, IV-62
 - randomized block designs, IV-19
 - t-tests, IV-19
 - two-sample t-test, IV-53, IV-72
 - two-sample z-test, IV-48
 - two-way ANOVA, IV-26, IV-57, IV-80
 - usage, IV-62
 - z-tests, IV-19
- power curves, IV-62
 - overlying curves, IV-67
 - response surfaces, IV-67
- Power model, IV-391, IV-405
- prediction intervals, II-40, II-46
- preference curves, IV-4
- preference mapping, IV-2
- PREFMAP, IV-7
- PRESS statistic
 - in partial least squares regression, III-360
- principal components, I-463
- principal components analysis
 - coefficients, I-456
 - compared to factor analysis, I-460
 - compared to linear regression, I-455
 - loadings, I-456
- prior probabilities, I-398
- probability calculator
 - examples, IV-90, IV-93, IV-94, IV-95
 - overview, IV-85
 - usage, IV-90
- probability limits, IV-121
- probability plots, I-15, II-9
- probit analysis
 - AIC and Schwarz's BIC, IV-99
 - algorithms, IV-107
 - categorical variables, IV-102
 - commands, IV-103
 - data format, IV-103
 - dummy coding, IV-102
 - effect coding, IV-103
 - examples, IV-104, IV-106
 - interpretation, IV-100
 - missing data, IV-107
 - model, IV-100
 - overview, IV-99
 - Quick Graphs, IV-103
 - saving files, IV-103
 - usage, IV-103
- process capability analysis, IV-155
 - Box-Cox power transformation, IV-157
 - non-normal data, IV-157, IV-158
 - process performance, IV-158
- Procrustes rotations, IV-7
- proportional hazards models, IV-433
- proportions
 - power analysis, IV-19, IV-25, IV-39, IV-40,

- IV-63
- p-value, IV-20
- Q
- QSK
- coefficients, I-172
- Quade test, III-329
- multiple comparisons, III-329
 - pairwise comparisons, III-330
- quadrat counts, IV-385, IV-398
- quadratic contrasts, II-28
- quality analysis, IV-109
- aggregated data, IV-120
 - average run length curves, IV-136
 - Box-and-Whisker plots, IV-112
 - commands, IV-161
 - control charts, IV-114
 - control limits, IV-121
 - cusum charts, IV-142
 - data format, IV-162
 - discrete control limits, IV-121
 - examples, IV-163, IV-164, IV-165, IV-166, IV-167, IV-168, IV-176, IV-178, IV-180, IV-183, IV-189, IV-191, IV-195, IV-197, IV-198, IV-199, IV-201, IV-203, IV-204, IV-206, IV-207, IV-209, IV-212, IV-213, IV-215
 - histogram, IV-110
 - moving average chart, IV-144
 - moving range, IV-149
 - operating characteristic curves, IV-135
 - overview, IV-109
 - Pareto charts, IV-111
 - process capability analysis, IV-155
 - quick graphs, IV-162
 - raw data, IV-120
 - regression charts, IV-152
 - run charts, IV-114
 - run tests, IV-118
 - shewhart control charts, IV-116
 - sigma limits, IV-122
 - TSQ charts, IV-153
 - usage, IV-162
 - X-MR charts, IV-149
- quantile plots, IV-434
- quantitative symmetric dissimilarity coefficient, I-162
- quartimax rotation, I-460, I-464
- quasi-independence, III-98
- Quasi-Newton method, III-269, III-273
- R
- R charts, IV-128
- R charts:plotting with X-bar charts, IV-129
- R matrix, II-289
- Ramsay procedure, III-279
- random coefficient models
- see mixed regression
- random effects, II-259, II-390
- in mixed regression, II-421
- random fields, IV-386
- random samples, I-8
- random sampling
- algorithms, IV-228
 - commands, IV-223
 - examples, IV-225, IV-226
 - overview
 - Quick Graphs, IV-224
 - univariate continuous, IV-222
 - univariate discrete, IV-220
 - usage, IV-224
- random variables, II-6
- random walk, IV-517
- randomized block designs, IV-37
- power analysis, IV-19
- range, I-301, I-307, IV-392
- Rank, IV-262
- rank regression, IV-262
- rank-order coefficients, I-172
- Rasch model, IV-490
- receiver operating characteristic curves
- See signal detection analysis
- regression

- bayesian regression, II-50
- LAD regression, IV-260
- Least-squares regression, IV-256
- linear, I-11
- LMS regression, IV-261
- logistic, III-1
- LTS regression, IV-261
- M-regression, IV-261
- rank regression, IV-262
- ridge regression, II-48
- S regression, IV-262
- TSLs regression, IV-581
- two-stage least squares, IV-581
- regression charts, IV-152
- regression trees, I-45
 - algorithms, I-62
 - basic tree model, I-42
 - commands, I-54
 - compared to analysis of variance, I-45
 - compared to stepwise regression, I-46
 - data format, I-54
 - displays, I-51
 - examples, I-55, I-57, I-59
 - loss functions, I-48, I-51
 - missing data, I-62
 - mobiles, I-41
 - model, I-51
 - overview, I-41
 - pruning, I-47
 - Quick Graphs, I-54
 - resampling, I-41
 - saving files, I-54
 - stopping criteria, I-47, I-53
 - usage, I-54
- R-E-G-W Q test, II-197
- R-E-G-W-Q test, II-27, II-119
- reliabilities, IV-492
- reliability, IV-489
- repeated measures, II-31
 - assumptions, II-32
- resampling
 - algorithms, I-38
 - bootstrap-t method, I-19
 - command, I-22
 - examples, I-23, I-27, I-28, I-33, I-34, I-36
 - missing data, I-38
 - naive bootstrap, I-19
 - overview, I-17
 - Quick Graphs, I-22
 - usage, I-22
- response optimization, IV-234
 - canonical analysis, IV-234
 - desirability analysis, IV-236
 - ridge analysis, IV-235
- response surface designs, I-350, I-354
 - analysis of, I-357
 - Box-Behnken designs, I-357
 - central composite designs, I-356
 - examples, I-380, I-384
 - rotatability, I-355, I-356
- response surface methods, IV-231
 - commands, IV-244
 - contour and surface plot, IV-233, IV-243
 - customization, IV-238
 - estimate model, IV-237, IV-238
 - examples, IV-245, IV-247, IV-249, IV-250
 - lack of fit, IV-233
 - optimize, IV-240
 - overview, IV-231
 - Quick Graphs, IV-244
 - usage, IV-244
- response surfaces, I-132, III-273
- restricted/residual maximum likelihood estimates, II-385
- ridge regression, II-48
- right censored data, IV-428
- RMSEA, III-457
- robust discriminant analysis, I-399
- robust regression
 - commands, IV-279
 - examples, IV-280, IV-283, IV-284
 - LAD regression, IV-260
 - LMS regression, IV-261
 - LTS regression, IV-261
 - M-regression, IV-261
 - overview, IV-255

- Quick Graphs, IV-279
- rank regression, IV-262
- S regression, IV-262
 - usage, IV-279
- robust smoothing, IV-358, IV-365
- robustness, III-321
- ROC curves, IV-320
- root mean square error of approximation, III-457
- rotatability
 - in response surface designs, I-355
- rotatable designs
 - in response surface designs, I-356
- rotation, I-459
- Roy's Greatest root, III-226
- running median smoothers, IV-512
- running-means smoother, IV-360
- S
 - s charts, IV-126
 - plotting with X-bar charts, IV-129
 - Sakitt D, IV-321
 - sample size, IV-23, IV-30
 - samples, I-8
 - saturated models
 - loglinear modeling, III-95
 - scale regression, IV-262
 - scalogram
 - see* partially ordered scalogram analysis with coordinates
 - scatterplot matrix, I-160
 - Scheffé model
 - in mixture designs, I-361
 - Scheffé test, II-27, II-118, II-197, II-307, II-395
 - screening designs, I-360
 - SD-RATIO, IV-321
 - seasonal decomposition, IV-523
 - second-order stationarity, IV-387
 - semi-variograms, IV-388
 - set correlations
 - assumptions, IV-292
 - categorical variables, IV-301
 - data format, IV-304
 - measures of association, IV-293
 - missing data, IV-316
 - overview, IV-291
 - partialing, IV-292
 - usage, IV-304
 - Shapiro-Wilk test, I-302
 - Shepard diagrams, III-189, III-194
 - Shepard's smoother, IV-360
 - Shewhart control charts
 - c charts, IV-131
 - np charts, IV-129
 - p charts, IV-130
 - R charts, IV-128
 - s charts, IV-126
 - u charts, IV-133
 - variance charts, IV-124
 - X charts, IV-129
 - X-bar charts, IV-123
 - Sidak test, II-27, II-118, II-197, II-307, II-395
 - sign test, III-325, III-326
 - signal detection analysis
 - algorithms, IV-346
 - chi-square model, IV-323
 - commands, IV-324
 - convergence, IV-324
 - data format, IV-325
 - examples, IV-328, IV-333, IV-335, IV-336, IV-340, IV-342, IV-344
 - exponential model, IV-323
 - gamma model, IV-323
 - logistic model, IV-323
 - missing data, IV-346
 - nonparametric model, IV-323
 - normal model, IV-323
 - overview, IV-319
 - poisson model, IV-323
 - Quick Graphs, IV-327
 - ROC curves, IV-327
 - usage, IV-325
 - sill, IV-392
 - similarity measures, I-157
 - simple matching dichotomy coefficients, I-164, I-173

- simplex, I-359
- Simplex method, III-269, III-273
- SIMPLS (Straight-forward IMplementation of Partial Least Squares)
 - see partial least squares regression
 - , III-377
- simulation, IV-394
- singular value decomposition, I-201, IV-6, IV-16
- skewness, I-307
 - positive, I-4
- slope, II-13
- smoothing, IV-362, IV-510
 - bandwidth, IV-350, IV-355
 - biweight kernel, IV-362, IV-364, IV-365
 - Cauchy kernel, IV-362, IV-365
 - commands, IV-366
 - confidence intervals, IV-368
 - data format, IV-366
 - discontinuities, IV-360
 - discrete gaussian convolution, IV-361
 - distance-weighted least squares (DWLS), IV-361
 - Epanechnikov kernel, IV-362, IV-364
 - examples, IV-367, IV-368, IV-370, IV-380
 - fixed-bandwidth method, IV-355, IV-362, IV-364
 - Gaussian kernel, IV-362, IV-364, IV-365
 - grid points, IV-361, IV-362, IV-382
 - inverse-distance, IV-360
 - k nearest-neighbors method, IV-356
 - kernel functions, IV-350, IV-352, IV-362, IV-364
 - LOESS smoothing, IV-361, IV-362, IV-367, IV-368, IV-370, IV-380
 - Marron & Nolan canonical kernel width, IV-357, IV-362, IV-364
 - mean smoothing, IV-358, IV-365
 - median smoothing, IV-358
 - methods, IV-350, IV-358, IV-365
 - model, IV-362
 - moving-averages, IV-360
 - Nadaraya-Watson, IV-360
 - nonparametric vs. parametric, IV-350
 - overview, IV-349
 - polynomial smoothing, IV-358, IV-365
 - Quick Graphs, IV-366
 - resampling, IV-349
 - residuals, IV-362, IV-366
 - robust smoothing, IV-358, IV-365
 - running-means, IV-360
 - saving results, IV-364, IV-366, IV-367
 - Shepard's smoother, IV-360
 - step, IV-361
 - tied values, IV-361
 - tricube kernel, IV-364, IV-365
 - trimmed mean smoothing, IV-365
 - triweight kernel, IV-364, IV-365
 - uniform kernel, IV-364
 - usage, IV-366
 - window normalization, IV-357, IV-364
- Sneath and Sokal's binary similarity coefficient, I-164
- Somers' d coefficients, I-227, I-235
- Sorting, I-5
- spaghetti plot, II-458
- spatial statistics, IV-385
 - algorithms, IV-426
 - azimuth, IV-403
 - commands, IV-408
 - data, IV-410
 - dip, IV-403
 - examples, IV-411, IV-417, IV-418, IV-424
 - grid, IV-407
 - kriging, IV-393, IV-400, IV-405
 - lags, IV-402
 - missing data, IV-426
 - model, IV-385, IV-403
 - nested models, IV-392
 - nesting structures, IV-403
 - nugget, IV-392
 - nugget effect, IV-392, IV-405
 - plots, IV-401
 - point statistics, IV-400
 - Quick Graphs, IV-410
 - resampling, IV-385
 - sill, IV-405

- simulation, IV-394, IV-401
- spherical model, IV-404
- trends, IV-406
- usage, IV-410
- variogram, IV-400
- Spearman coefficients, I-162, I-172, I-227
- Spearman-Brown coefficient, IV-489
- specificities, I-458
- spectral models, IV-510
- spherical model, IV-389
- split plot designs, II-175
- split-half reliabilities, IV-492
- SSCP matrix, III-135
- standard deviation, I-3, I-301, I-307
- standard error of estimate, II-7
- standard error of skewness, I-307
- standard error of the mean, I-11, I-307
- standardization, I-67
- standardized alpha, IV-489
- standardized deviates, I-202
- standardized values, I-6
- stationarity, IV-387, IV-520
- statistics
 - defined, I-1
 - descriptive, I-1
 - inferential, I-7
- stem-and-leaf plots, I-3, I-299
- step smoother, IV-361
- stepwise regression, II-15, II-30, III-9
- stochastic processes, IV-386
- stress, III-188, III-211
- structural equation models
 - see path analysis
- Stuart's tau-c coefficients, I-227, I-234
- Student, II-197
- studentized residuals, II-10
- Student-Newman-Keuls test, II-27, II-119
- subpopulations, I-305
- subsampling, I-18
- sum of cross-products matrix, I-171
- sums of squares
 - type I, II-29, II-34, II-113
 - type II, II-35, II-113
 - type III, II-30, II-36, II-113
 - type IV, II-36
- surface plot, IV-243
- surface plots, IV-401
- survival analysis
 - AIC and Schwarz's BIC, IV-427
 - algorithms, IV-476
 - censoring, IV-428, IV-435, IV-479
 - centering, IV-477
 - coding variables, IV-437
 - commands, IV-447
 - convergence, IV-481
 - Cox regression, IV-448
 - data format, IV-448
 - estimation, IV-442
 - examples, IV-449, IV-453, IV-455, IV-459, IV-462, IV-464, IV-468, IV-472
 - exponential model, IV-441
 - graphs, IV-437, IV-444
 - logistic model, IV-441
 - log-likelihood, IV-477
 - lognormal model, IV-435, IV-477
 - missing data, IV-476
 - model, IV-435
 - models, IV-479
 - Nelson-Aalen cumulative hazard estimator, IV-438
 - overview, IV-427
 - parameters, IV-476
 - plots, IV-481
 - proportional hazards models, IV-479
 - Quick Graphs, IV-448
 - Singular Hessian, IV-478
 - stepwise, IV-482
 - stepwise estimation, IV-443
 - tables, IV-437, IV-444
 - time dependent covariates, IV-446
 - usage, IV-448
 - variances, IV-483
 - weibull model, IV-472
- symmetric matrix, I-160

T

- t tests
- Taguchi designs, I-353, I-377
- Tamhane's T2 test, II-27, II-119, II-197
- Tanimoto dichotomy coefficients, I-164, I-173
- tau-b coefficients, I-234
- tau-c coefficients, I-234
- test for normality, I-302
 - Anderson-Darling test, I-303
 - Shapiro-Wilk test, I-302
- test item analysis
 - algorithms, IV-506
 - classical analysis, IV-488, IV-489, IV-491, IV-506
 - commands, IV-494
 - data format, IV-495
 - examples, IV-498, IV-500, IV-503
 - logistic item-response analysis, IV-490, IV-493, IV-506
 - missing data, IV-507
 - overview, IV-487
 - Quick Graphs, IV-497
 - reliabilities, IV-492
 - resampling, IV-487
 - scoring items, IV-492, IV-493
 - statistics, IV-495
 - usage, IV-495
- tests for correlation, I-535
 - equality of two correlations, I-522, I-537
 - specific correlation, I-522, I-536
 - zero correlation, I-522, I-535
- tests for mean, I-523
 - one-sample t, I-520, I-526
 - one-sample z, I-520, I-523
 - paired t, I-521, I-527
 - poisson, I-520, I-530
 - two-sample t, I-521, I-528
 - two-sample z, I-520, I-524
- tests for normality
 - AD test, III-334
 - K-S test, III-331
 - Lilliefors test, III-334
 - Shapiro-Wilk's test, I-497
- tests for proportion, I-538
 - equality of proportions, I-521
 - equality of two proportions, I-540
 - single proportion, I-520, I-538
- tests for variance, I-531
 - Bartlett's test, I-521
 - equality of several variances, I-534
 - equality of two variances, I-521, I-532
 - Levene's test, I-521
 - single variance, I-531
- tetrachoric correlation, I-164, I-166
- theory of signal detectability (TSD), IV-319
- time domain models, IV-510
- time series, IV-509
 - algorithms, IV-578
 - ARIMA models, IV-514, IV-540
 - clear series, IV-534
 - commands, IV-532, IV-534, IV-539, IV-540, IV-542, IV-544, IV-546
 - data format, IV-546
 - examples, IV-547, IV-548, IV-549, IV-550, IV-552, IV-555, IV-557, IV-558, IV-560, IV-561, IV-566, IV-575
 - forecasts, IV-538
 - Fourier transformations, IV-545
 - missing values, IV-509
 - moving average, IV-511, IV-535
 - overview, IV-509
 - plot labels, IV-528
 - plots, IV-528, IV-529, IV-530, IV-531
 - Quick Graphs, IV-546
 - running means, IV-512, IV-535
 - running medians, IV-512, IV-536
 - seasonal adjustments, IV-523, IV-539
 - smoothing, IV-510, IV-535, IV-536, IV-537
 - stationarity, IV-520
 - transformations, IV-532, IV-534
 - trend analysis, IV-525, IV-542
 - trends, IV-538
 - usage, IV-546
- tolerance, II-16
- T-plots, IV-529

- trace criterion
 - see A-optimality
- tree clustering methods, I-47
- tree diagrams, I-70
- trend analysis, IV-525, IV-542
 - Homogeneity test, IV-544
 - Mann-Kendall test, IV-526, IV-543
 - Modified Seasonal Kendall test, IV-543
 - Seasonal Kendall test, IV-526, IV-543
 - slope estimator, IV-573
- triangle inequality, III-186
- tricube kernel, IV-364
- trimmed mean, I-299, I-308
- trimmed mean smoothing, IV-365
- triweight kernel, IV-364
- t-tests, IV-19
 - one-sample, I-526, IV-50
 - paired, I-527, IV-51
 - power analysis, IV-26
 - two-sample, I-528, IV-53
- Tukey procedure, III-279
- Tukey test, II-27, II-118, II-196
- Tukey's b test, II-27, II-119, II-197
- Tukey's HSD test, II-307, II-395
- Tukey's jackknife, I-18
- twoing, I-48
- two-stage least squares
 - algorithms, IV-597
 - commands, IV-586
 - estimation, IV-582
 - examples, IV-587, IV-590, IV-592, IV-593, IV-595, IV-596
 - heteroskedasticity-consistent standard errors, IV-586
 - lagged variables, IV-586
 - missing data, IV-597
 - model, IV-585
 - overview, IV-581
 - Quick Graphs, IV-586
 - usage, IV-586
- Type I error, IV-21
- Type II error, IV-22
- U
 - u charts, IV-133, IV-134
 - unbalanced designs
 - in analysis of variance, II-29
 - uncertainty coefficient, I-234
 - unfolding models, IV-3
 - uniform kernel, IV-364
- V
 - validity, I-87
 - variance, I-307
 - of estimates, I-355
 - variance charts, IV-124
 - variance component models
 - see mixed regression
 - variance components
 - categorical variables, II-303
 - commands, II-310
 - examples, II-311, II-315, II-320, II-323, II-326, II-328, II-334, II-340
 - hypothesis test, II-306
 - model estimation, II-301
 - models, II-301
 - options, II-304
 - overview, II-299
 - Quick Graph, II-310
 - usage, II-310
 - variance inflation factor, II-70
 - variance of prediction, I-356
 - variance paths
 - path analysis, III-401
 - varimax rotation, I-460, I-464
 - variograms, IV-388, IV-401
 - model, IV-389
 - vector model
 - in perceptual mapping, IV-5
 - Voronoi polygons, IV-385, IV-397, IV-400
- W
 - Wald-Wolfowitz runs test, III-337
 - wave model, IV-391

Weibull, III-334
Weibull distribution, IV-432
weighted running smoothing, IV-512
weights, I-23, I-54, I-135, I-179, I-206, I-246, I-248, I-323, I-371, I-408, I-469, I-503, I-544, II-54, II-121, II-122, II-202, II-311, II-357, II-399, II-441, II-442, III-23, III-103, III-104, III-137, III-194, III-217, III-283, III-339, III-340, III-364, III-385, III-413, IV-9, IV-63, IV-104, IV-162, IV-244, IV-280, IV-305, IV-325, IV-328, IV-366, IV-367, IV-410, IV-449, IV-495, IV-498, IV-547, IV-587
Wilcoxon Signed-Rank test, III-326
Wilcoxon test, III-326
Wilk's trace, I-405
Wilks' lambda, I-405, III-225
Winter's three-parameter model, IV-524
Within-Group Testing, III-241, III-257
within-subjects differences
 in analysis of variance, II-32

X

X charts, IV-129
X-bar charts, IV-123
 plotting with R charts, IV-129
 plotting with s charts, IV-129
X-MR charts, IV-149
 control limits, IV-149

Y

Yates' correction, I-226, I-233
y-intercept, II-12
Young's S-STRESS, III-190
Yule's Q, I-228
Yule's Q coefficient, I-164
Yule's Y, I-228, I-234

Z

z tests
z-tests, IV-19
 one-sample, IV-46

two-sample, IV-48